L'informazione come fondamento

Entropia, inferenza bayesiana e i modelli generativi dell'intelligenza artificiale

Introduzione — Il secondo paradigma e il tramonto del simbolo

Arriva un momento, nella storia intellettuale di una disciplina, in cui la sostituzione di un paradigma con un altro non si manifesta come semplice revisione tecnica ma come capovolgimento dell'intera struttura concettuale che sosteneva il precedente edificio. L'intelligenza artificiale ha attraversato esattamente questa trasformazione nel corso degli ultimi decenni, e la comprensione profonda di ciò che è accaduto richiede non soltanto la descrizione dei nuovi strumenti matematici adottati, ma l'analisi delle ragioni per cui i vecchi strumenti erano costitutivamente inadeguati — non per difetti contingenti di implementazione, bensì per insufficienze strutturali radicate nella loro natura più intima. Questa diagnosi richiede però una cautela metodologica che sarà mantenuta lungo tutto il saggio: il confronto tra paradigmi non può essere condotto scegliendo come bersaglio la versione più vulnerabile del paradigma soppiantato. La storia intellettuale esige che si affrontino le formulazioni più raffinate, non le più facili da confutare.

Silvano Salvador

Il paradigma che ha dominato la ricerca sull'intelligenza artificiale dalla metà degli anni cinquanta fino all'inizio degli anni novanta era fondato su una premessa apparentemente solida: l'intelligenza è essenzialmente manipolazione di simboli secondo regole esplicite. Questa ipotesi, che Newell e Simon avevano elevato a principio metodologico nella loro celebre tesi sui sistemi fisici di simboli del 1976, non era arbitraria né priva di motivazioni filosofiche profonde. Essa si inseriva in una tradizione secolare che risaliva almeno a Leibniz e alla sua visione di una characteristica universalis — un linguaggio simbolico perfetto nel quale ogni problema intellettuale potesse essere ridotto a calcolo meccanico. La logica matematica del XX secolo, da Frege a Russell, da Hilbert a Gödel, sembrava aver fornito la struttura formale adeguata per realizzare questa visione.

Il collasso di questa prospettiva fu duplice. Da un lato i teoremi di incompletezza di Gödel e il risultato di indecidibilità di Turing dimostrarono che nessun sistema formale sufficientemente espressivo può essere al contempo completo e consistente, e che non esiste alcun procedimento algoritmico generale capace di decidere la verità aritmetica. Dall'altro, sul piano applicativo, l'intelligenza artificiale simbolica si arenò su difficoltà sistematiche: il problema del frame, la fragilità nei domini aperti, l'impossibilità di codificare la conoscenza di senso comune. Questi due fallimenti — l'uno matematico, l'altro ingegneristico — convergevano verso la medesima diagnosi: la logica formale non è il fondamento dell'intelligenza, ma uno dei suoi prodotti più elevati e, al tempo stesso, uno strumento di portata inevitabilmente limitata quando applicato a domini naturalmente aperti e statisticamente strutturati.

Occorre però segnalare immediatamente ciò che questo saggio non intende sostenere. La diagnosi appena formulata non equivale ad affermare che il paradigma simbolico abbia fallito in ogni suo aspetto o

Silvano Salvador

che le sue linee di sviluppo successive siano prive di valore. Le description logics — sviluppate sistematicamente da Brachman e Levesque negli anni ottanta e formalizzate nel Web Ontology Language — non soffrono del frame problem perché operano su world descriptions esplicitamente chiuse, una scelta epistemologica deliberata che riflette i requisiti di domini applicativi precisi. I model checkers come SPIN e NuSMV verificano proprietà di sistemi concorrenti con milioni di configurazioni producendo garanzie che nessun sistema statistico può offrire. I proof assistants — Coq, Isabelle/HOL, Lean 4 — permettono la formalizzazione e la verifica meccanica di dimostrazioni matematiche di complessità straordinaria: la dimostrazione del teorema di Feit-Thompson in Coq ha richiesto oltre 150.000 righe di codice e sei anni di lavoro, producendo una certezza epistemologicamente distinta da qualsiasi predizione probabilistica. In avionica, nella progettazione di processori, nella sicurezza crittografica, la verifica simbolica formale rimane indispensabile e insostituibile.

Il confronto tra paradigmi che questo saggio propone non è dunque quello tra una tradizione fallita e una vittoriosa. È il confronto tra due architetture concettuali con domini di applicazione distinti, e con strutture matematiche adeguate a classi di problemi diverse. Il paradigma simbolico eccelle nei domini chiusi, formalmente specificabili, dove la correttezza deve essere certificata e la catena inferenziale deve essere ispezionabile. Il paradigma informazionale — quello che questo saggio si propone di esporre e analizzare — eccelle nei domini aperti, statisticamente strutturati, dove la generalizzazione da esempi parziali è necessaria e l'incertezza deve essere quantificata e propagata. La domanda rilevante non è quale dei due paradigmi sia corretto in assoluto, ma in quali condizioni ciascuno offre i migliori compromessi tra le proprietà desiderabili di un sistema intelligente.

Silvano Salvador

Il presente saggio muove da questo punto di partenza e si propone di mostrare quale sia la struttura matematica che ha effettivamente permesso lo sviluppo dell'intelligenza artificiale moderna nei domini aperti, e perché essa risulti non soltanto tecnicamente più efficace del paradigma simbolico per quella classe di problemi, ma concettualmente più adeguata alla natura statistica dei fenomeni cognitivi che essi richiedono di modellare. Quella struttura è la teoria dell'informazione, intesa nel senso ampio che abbraccia i contributi di Shannon, Kolmogorov, Jaynes, Rissanen e la tradizione bayesiana nella sua totalità. Non si tratta di un insieme eterogeneo di tecniche, bensì di un'unica architettura concettuale che si dispiega dalla definizione elementare di entropia fino ai modelli generativi profondi contemporanei, passando per la termodinamica statistica, l'inferenza variazionale e la geometria delle distribuzioni di probabilità.

La tesi che si intende sostenere è la seguente: l'informazione — nel senso tecnico e preciso che le attribuiranno le pagine seguenti — non è un concetto ausiliare da affiancare ad altri strumenti matematici, ma la struttura profonda che unifica apprendimento, compressione, percezione e generazione in un'unica teoria. I modelli generativi che oggi costituiscono lo stato dell'arte nell'intelligenza artificiale — dai modelli di diffusione stocastica ai transformer autoregressivi, dai modelli a energia alle reti variazionali — sono implementazioni computazionali di principi inferenziali che si trovano già formulati, con straordinaria chiarezza, nella termodinamica statistica del XIX secolo e nella teoria della comunicazione del XX. Riconoscere questa continuità non significa sminuire la novità delle architetture recenti, ma comprenderne le fondamenta con una profondità che la presentazione tecnica isolata non consente.

Silvano Salvador

Rispetto al paradigma simbolico nei domini cognitivi aperti, l'approccio informazionale comporta un rovesciamento metodologico radicale. Laddove la logica simbolica procede per regole esplicite, strutture di dati dichiarative e inferenze deterministiche, la teoria dell'informazione opera su distribuzioni di probabilità, incertezza quantificata e aggiornamenti bayesiani. Laddove i sistemi esperti richiedevano la specifica manuale di ogni relazione tra concetti, i sistemi neurali profondi apprendono rappresentazioni latenti da distribuzioni empiriche di dati. Questa non è semplicemente una differenza di efficienza computazionale: è una differenza nell'ontologia del problema che si ritiene di star risolvendo. L'intelligenza simbolica assumeva che il mondo si lasciasse decomporre in entità discrete connesse da relazioni esplicitabili; l'intelligenza statistica assume invece che il mondo sia fondamentalmente una distribuzione di probabilità su configurazioni possibili, e che comprendere significhi saper approssimare tale distribuzione con sufficiente accuratezza.

Parte Prima — La misura dell'incertezza: Shannon e la geometria dell'informazione

I. La nascita di una teoria

Nell'estate del 1948 Claude Shannon pubblicò sulla Bell System Technical Journal un articolo intitolato A Mathematical Theory of Communication, destinato a trasformare la comprensione stessa di ciò che significa trasmettere, ricevere e processare informazione. La portata di quell'articolo eccedeva largamente i problemi ingegneristici della trasmissione telefonica che ne avevano motivato la stesura: Shannon stava fondando una scienza dell'informazione in senso assoluto,

Silvano Salvador

indipendente dal contenuto semantico dei messaggi e dalla natura fisica del canale di trasmissione.

Il punto di partenza era una domanda apparentemente semplice: quanto è incerta una sorgente di messaggi? Se una sorgente discreta emette simboli da un alfabeto finito $\mathscr{X} = \{x_1, x_2, \ldots, x\}$ con probabilità rispettive $p_1, p_2, \ldots, p$, quale numero reale misura in modo appropriato la nostra incertezza prima di osservare quale simbolo verrà emesso? Shannon dimostrò che, imponendo tre condizioni naturali — continuità rispetto alle probabilità, simmetria rispetto alla permutazione dei simboli equiprobabili, e coerenza con il calcolo ricorsivo dell'incertezza composita — l'unica funzione che le soddisfa tutte è, a meno di una costante positiva di scala:

$$H(p_1, p_2, \ldots, p) = -\sum_i p_i \log p_i$$

dove la convenzione più comune in teoria dell'apprendimento adotta il logaritmo naturale, mentre la convenzione dell'ingegneria delle comunicazioni impiega il logaritmo in base 2, misurando l'entropia in bit. Shannon chiamò questa grandezza entropia, in analogia esplicita — suggerita da von Neumann — con la magnitudine termodinamica omonima di Boltzmann e Gibbs.

L'entropia H possiede proprietà che meritano di essere esplorate con attenzione, perché esse riflettono strutture concettuali di grande rilevanza per la teoria dell'intelligenza. $H \geq 0$, con uguaglianza se e soltanto se una delle probabilità è uguale a uno: l'incertezza è nulla quando il risultato è già determinato. H raggiunge il proprio massimo — uguale a $\log n$ — quando la distribuzione è uniforme: l'incertezza è massima quando tutte le possibilità sono equiprobabili, quando nessuna struttura discriminante è presente nella sorgente. L'informazione, in questo senso tecnico, è ciò che riduce l'incertezza: apprendere il valore di

Silvano Salvador

una variabile con distribuzione p significa acquisire H(p) bit di informazione in media.

Il legame tra entropia e compressione ottimale è stabilito dal teorema di codifica della sorgente: il numero medio di bit necessari per codificare i simboli di una sorgente con distribuzione p è almeno H(p), e codici opportunamente costruiti permettono di avvicinarsi a questo limite con arbitraria precisione. L'entropia è dunque il limite fondamentale della compressione senza perdita. Questa connessione — tra misura dell'incertezza e compressione ottimale — anticipa il tema sviluppato nella Parte Quarta per cui intelligenza e compressione sono, in un senso matematicamente preciso, la medesima operazione eseguita su scale differenti.

II. L'entropia condizionale e l'informazione mutua

La generalizzazione più naturale dell'entropia a coppie di variabili casuali introduce due grandezze fondamentali. Se X e Y sono variabili casuali con distribuzione congiunta p(x, y), l'entropia condizionale di Y dato X è:

$$H(Y|X) = -\sum_{x} p(x) \sum_{y} p(y|x) \log p(y|x) = -\sum_{x} \sum_{y} p(x, y) \log p(y|x)$$

e rappresenta l'incertezza residua su Y dopo aver osservato X. Si verifica che $H(Y|X) \leq H(Y)$, con uguaglianza se e soltanto se X e Y sono statisticamente indipendenti: osservare una variabile non può, in media, aumentare la nostra incertezza su un'altra.

L'informazione mutua tra X e Y è:

$$I(X; Y) = H(Y) - H(Y|X) = H(X) - H(X|Y) = H(X) + H(Y) - H(X, Y)$$

Silvano Salvador

L'informazione mutua misura quanto la conoscenza di una variabile riduce l'incertezza sull'altra, ed è simmetrica — $I(X; Y) = I(Y; X)$ — e sempre non negativa, con $I(X; Y) = 0$ se e soltanto se X e Y sono indipendenti. Nelle reti neurali profonde l'informazione mutua tra rappresentazioni interne e variabili di input o output ha assunto un ruolo importante nell'analisi teorica dell'apprendimento, in particolare nel quadro dell'ipotesi del collo di bottiglia dell'informazione di Tishby, Pereira e Bialek. Va però segnalato che gli sviluppi empirici di questa linea di ricerca hanno mostrato che le previsioni originali richiedono significative qualificazioni: la fase di compressione che Tishby e Zaslavsky (2015) avevano ipotizzato durante l'addestramento non si osserva nelle reti con attivazioni deterministiche come ReLU, ma soltanto con attivazioni che introducono stocasticità esplicita. La questione del collo di bottiglia sarà ripresa in dettaglio nella Parte Quarta.

III. La divergenza di Kullback-Leibler

La grandezza più centrale per la teoria dell'apprendimento automatico — quella che unifica la verosimiglianza massima, la perdita per entropia crociata e l'inferenza variazionale in un'unica struttura — non è l'entropia in sé ma la divergenza di Kullback-Leibler, introdotta da Solomon Kullback e Richard Leibler nel 1951. Se P e Q sono due distribuzioni di probabilità su uno spazio misurabile $\mathcal{X}$, la divergenza da P a Q è:

$$D_{KL}(P \parallel Q) = \sum_{x} p(x) \log \left[p(x) / q(x) \right]$$

nella versione discreta, o

$$D_{KL}(P \parallel Q) = \int p(x) \log \left[p(x) / q(x) \right] d\mu(x)$$

nella versione continua rispetto a una misura dominante μ. Per convenzione, si pone $0 \cdot \log(0/q) = 0$ e $p \cdot \log(p/0) = +\infty$ qualora $p > 0$: la divergenza è infinita quando Q non domina P.

Silvano Salvador

La divergenza KL è sempre non negativa — $D_{KL}(P \| Q) \geq 0$, con uguaglianza se e soltanto se $P = Q$ quasi ovunque — come conseguenza immediata della disuguaglianza di Jensen applicata alla funzione convessa $-\log$. Essa non è simmetrica in generale: $D_{KL}(P \| Q) \neq D_{KL}(Q \| P)$, e non soddisfa la disuguaglianza triangolare, per cui non è una distanza nel senso metrico ordinario. Queste asimmetrie non sono difetti tecnici ma riflettono strutture concettuali genuine di grande rilevanza pratica.

La divergenza $D_{KL}(Q \| P)$ — letta nella direzione opposta — ha un comportamento qualitativamente diverso da $D_{KL}(P \| Q)$. La prima è detta zero-forcing: quando si minimizza $D_{KL}(Q \| P)$ rispetto a Q, la distribuzione approssimante tende a concentrarsi sulle mode di P, trascurando le regioni a bassa densità anche quando esse contengono massa significativa, poiché $q(x) \to 0$ per x in regioni dove $p(x) > 0$ non produce penalità infinite. La seconda, $D_{KL}(P \| Q)$, è mass-covering: se $p(x) > 0$ mentre $q(x) \to 0$, il termine $p(x) \log[p(x)/q(x)]$ diverge, per cui Q è penalizzata per assegnare densità nulla a regioni a cui P assegna massa positiva. Questa asimmetria ha conseguenze dirette nelle applicazioni di inferenza variazionale, dove la scelta della direzione della divergenza determina il tipo di approssimazione ottenuta — argomento che sarà sviluppato sistematicamente nella Parte Seconda.

Il legame con l'entropia crociata è immediato. L'entropia crociata di P rispetto a Q è:

$$H(P, Q) = -\sum_x p(x) \log q(x) = H(P) + D_{KL}(P \| Q)$$

Poiché $H(P)$ è fissa rispetto ai parametri del modello Q durante l'addestramento, minimizzare l'entropia crociata tra la distribuzione empirica dei dati e la distribuzione del modello è equivalente a minimizzare $D_{KL}(P \| Q)$, che equivale a sua volta alla massimizzazione della verosimiglianza logaritmica sui dati di addestramento. La pervasività

Silvano Salvador

della perdita per entropia crociata nell'apprendimento profondo non è dunque una scelta arbitraria ma la manifestazione di un principio unificante: addestrare un modello significa ridurre la divergenza informazionale tra la distribuzione del modello e la distribuzione generativa sottostante ai dati.

IV. Il principio di massima entropia

Il 1957 è un anno cruciale per la teoria dell'informazione applicata: Edwin Thompson Jaynes pubblicò due articoli su Physical Review nei quali mostrava che la meccanica statistica di Gibbs poteva essere rifondatata come applicazione di un principio inferenziale generale — il principio di massima entropia.

Il principio di Jaynes può essere enunciato come segue. Si abbia un insieme di vincoli lineari sulla distribuzione incognita p, della forma:

$$\sum_x p(x)\, f_k(x) = \mu_k, \quad k = 1, 2, \ldots, m$$

dove le f_k sono funzioni osservabili e le μ_k sono i loro valori attesi misurati o imposti. Tra tutte le distribuzioni di probabilità che soddisfano tali vincoli, il principio di massima entropia seleziona quella che massimizza $H(p)$. La soluzione si ottiene per mezzo dei moltiplicatori di Lagrange.

Il Lagrangiano è:

$$\mathcal{L}(p, \lambda, \lambda_0) = -\sum_x p(x) \log p(x) - \lambda_0 \left(\sum_x p(x) - 1\right) - \sum_k \lambda_k \left(\sum_x p(x) f_k(x) - \mu_k\right)$$

Differenziando rispetto a $p(x)$ e ponendo la derivata uguale a zero:

$$-\log p(x) - 1 - \lambda_0 - \sum_k \lambda_k f_k(x) = 0$$

da cui:

Silvano Salvador

$p^*(x) = \exp(-\sum \lambda \, f(x)) \,/\, Z(\lambda_1, \ldots, \lambda)$

dove $Z = \sum_x \exp(-\sum \lambda \, f(x))$ è la funzione di partizione normalizzante e i moltiplicatori λ sono determinati dal rispetto dei vincoli. La verifica che si tratti di un massimo globale deriva dalla concavità stretta di H: l'Hessiano della funzione obiettivo rispetto alle probabilità è $-\mathrm{diag}(1/p(x))$, definito negativo in tutto l'interno del simplesso.

La funzione di partizione Z svolge un ruolo generativo di straordinaria centralità: i valori attesi delle statistiche sufficienti f si ottengono come derivate logaritmiche di Z rispetto ai moltiplicatori:

$\mu = \mathbb{E}_\{p^*\}[f] = -\partial \log Z / \partial \lambda$

e la matrice di covarianza tra le statistiche è la matrice di curvatura di log Z:

$\mathrm{Cov}_\{p^*\}[f, f] = \partial^2 \log Z \,/\, (\partial \lambda \, \partial \lambda)$

che è definita positiva, poiché Z è log-convessa. Questa struttura è identica a quella della funzione cumulante di una distribuzione della famiglia esponenziale — le distribuzioni gaussiane, bernoulliane, poissoniane, gamma, dirichlet appartengono tutte a questa famiglia — il che spiega perché esse massimizzino l'entropia dati vincoli sui momenti. In altri termini, sono le distribuzioni meno presupposte compatibili con le informazioni disponibili: esse non introducono struttura aggiuntiva oltre a quella che i vincoli impongono.

L'importanza del principio di massima entropia per l'intelligenza artificiale è triplice. Sul piano teorico, esso fornisce una giustificazione assiomatica per l'uso di distribuzioni della famiglia esponenziale. Sul piano pratico, esso giustifica la regressione logistica, i modelli autoregressivi e le macchine di Boltzmann come architetture che implementano una forma di massimizzazione dell'entropia vincolata. Sul

Silvano Salvador

piano epistemologico, esso rappresenta la versione più generale del principio di parsimonia: non assumere più di quanto i dati costringano ad assumere.

V. La divergenza KL come distanza geometrica: l'informazione di Fisher

Il collegamento tra la divergenza di Kullback-Leibler e la metrica di Fisher costituisce il fondamento della geometria dell'informazione e la giustificazione teorica per alcune delle scelte più importanti nell'ottimizzazione di modelli probabilistici. Sia $\{p(x;\theta) : \theta \in \Theta\}$ una famiglia parametrica di distribuzioni, dove $\Theta \subseteq \mathbb{R}^d$ è aperto e $\theta \mapsto p(x;\theta)$ è sufficientemente regolare. La matrice di informazione di Fisher $G(\theta)$ è una matrice semidefinita positiva $d \times d$ con elementi:

$$G_{i}(\theta) = \mathbb{E}_{x \sim p(\cdot;\theta)}[\partial_i \log p(x;\theta) \cdot \partial \log p(x;\theta)]$$

Sotto condizioni di regolarità appropriate, questa espressione è equivalente a:

$$G_{i}(\theta) = -\mathbb{E}_{x \sim p(\cdot;\theta)}[\partial^2_{\theta_i \theta} \log p(x;\theta)]$$

che lega la matrice di Fisher alla curvatura della funzione di log-verosimiglianza. La connessione con la divergenza KL emerge dall'espansione al secondo ordine: la divergenza KL tra distribuzioni vicine nello spazio dei parametri è, al secondo ordine:

$$D_{KL}(p(\cdot;\theta) \,\|\, p(\cdot;\theta + d\theta)) = \tfrac{1}{2}\, d\theta^{T} G(\theta)\, d\theta + O(\|d\theta\|^3)$$

$G(\theta)$ definisce quindi una metrica riemanniana sullo spazio dei parametri — la metrica di Fisher — che conferisce alla famiglia di distribuzioni la struttura di una varietà differenziabile. Questa geometria è l'oggetto di studio della geometria dell'informazione, sviluppata sistematicamente da Shun'ichi Amari a partire dagli anni settanta.

Silvano Salvador

Il gradiente naturale, introdotto da Amari, è definito come:

$$\tilde{\nabla}_\theta L = G(\theta)^{-1} \nabla_\theta L$$

Mentre il gradiente ordinario indica la direzione di massima discesa nella metrica euclidea dello spazio dei parametri — metrica che dipende dalla parametrizzazione scelta e non ha interpretazione intrinseca — il gradiente naturale indica la direzione di massima discesa nella metrica di Fisher, invariante rispetto alla scelta della parametrizzazione. L'aggiornamento con gradiente naturale è invariante per reparametrizzazione: se si sostituisce θ con una funzione invertibile differenziabile $\eta = h(\theta)$, il gradiente naturale rispetto a η produce lo stesso aggiornamento nella distribuzione p che il gradiente naturale rispetto a θ. Il gradiente ordinario non gode di questa proprietà — la direzione di aggiornamento dipende dalla parametrizzazione — il che introduce un bias sistematico capace di rallentare la convergenza in presenza di parametri con scale molto diverse o con correlazioni forti.

Nell'ottimizzazione di modelli neurali profondi il gradiente naturale è computazionalmente proibitivo perché richiede l'inversione di una matrice $d \times d$ dove d è il numero di parametri, potenzialmente nell'ordine di miliardi. Tuttavia esso costituisce il fondamento teorico di approssimazioni praticamente importanti come K-FAC (Kronecker-factored Approximate Curvature), e la sua comprensione concettuale è indispensabile per apprezzare le dinamiche di addestramento. In particolare, il Natural Policy Gradient di Kakade (2001) e le sue varianti approssimate — TRPO e PPO — nel reinforcement learning sfruttano la stessa intuizione per stabilizzare e accelerare l'ottimizzazione della policy in spazi continui ad alta dimensionalità, tema che sarà approfondito nel terzo saggio che seguirà.

Silvano Salvador

V-bis. La geometria dell'informazione di Amari: varietà statistiche e connessioni duale

La metrica di Fisher definisce una geometria riemanniana sullo spazio statistico dei modelli probabilistici, ma la geometria dell'informazione — sviluppata sistematicamente da Shun-ichi Amari e Hiroshi Nagaoka a partire dagli anni Ottanta — rivela una struttura molto più ricca: lo spazio statistico è una varietà differenziabile equipaggiata non con una ma con una famiglia continua di connessioni, parametrizzate da un numero reale α.

La connessione α su una varietà statistica $M = \{p(x; \theta)\}$ è definita attraverso i simboli di Christoffel:

$$\Gamma^{(\alpha)}_{ij,k} = E[\partial_i \partial_j \log p \cdot \partial_k \log p] + (1-\alpha)/2 \cdot E[\partial_i \partial_j \partial_k \log p]$$

Le connessioni $\alpha = 1$ ed $\alpha = -1$ sono strutturalmente duali nel senso che una curva geodetica per la connessione $\alpha = 1$ (curva e-geodetica, o espettata) è il correlato duale di una curva geodetica per $\alpha = -1$ (curva m-geodetica, o mista). Questa dualità non è una semplice simmetria: ha conseguenze computazionali concrete. Il teorema di proiezione di Amari afferma che, dati uno spazio statistico E e una sottovarietà M, la proiezione e-geodetica di p su M coincide con la minimizzazione della divergenza KL $D_{KL}(M \| p)$ — ovvero con il problema dell'inferenza variazionale forward. La proiezione m-geodetica coincide con la minimizzazione di $D_{KL}(p \| M)$ — ovvero con la massima verosimiglianza.

L'importanza di questa dualità per l'apprendimento automatico è che rende precisa la relazione tra due famiglie di metodi che la pratica distingue intuitivamente ma raramente formalizza. L'inferenza variazionale — minimizzare $D_{KL}(q \| p)$ rispetto a q in una famiglia

Silvano Salvador

variazionale Q — è una proiezione e-geodetica di p su Q. La massima verosimiglianza — massimizzare log p(x; θ) rispetto a θ — è una proiezione m-geodetica della distribuzione empirica P_{emp} sulla famiglia parametrica. Questa unificazione geometrica spiega perché i due metodi hanno proprietà statistiche molto diverse: la proiezione e-geodetica ("zero-avoiding") evita di mettere massa su regioni dove p ha densità, producendo mode-seeking; la proiezione m-geodetica ("zero-forcing") media su tutte le mode della distribuzione target.

Per i modelli di diffusione, la geometria dell'informazione offre un'interpretazione elegante. Il processo di denoising è una traiettoria nella varietà statistica: parte dalla distribuzione gaussiana $p^T(x)$ − prossima alla distribuzione dell'informazione minima — e converge verso la distribuzione dei dati p_0(x) lungo le geodetiche della connessione $\alpha = -1$. La score function $\nabla_x \log p^T(x)$ è il vettore tangente alla traiettoria m-geodetica nello spazio delle distribuzioni, il che rende preciso perché l'apprendimento della score sia equivalente all'apprendimento della geometria locale della distribuzione dei dati.

Il contributo forse più duraturo della geometria dell'informazione di Amari all'apprendimento automatico è il gradiente naturale, discusso nella sezione precedente, visto ora come il gradiente nella metrica riemanniana della varietà statistica. La discesa del gradiente ordinaria è una traiettoria m-geodetica nello spazio dei parametri, ma ignora la curvatura della varietà: due passi della stessa lunghezza euclidea corrispondono a variazioni molto diverse della distribuzione p(x; θ) a seconda della curvatura locale. Il gradiente naturale corregge questa distorsione moltiplicando per l'inversa della matrice di Fisher $G(\theta)^{-1}$, producendo il gradiente nella metrica riemanniana dello spazio statistico. K-FAC (Martens e Grosse, 2015) e altri approssimatori del gradiente

Silvano Salvador

naturale per reti neurali profonde sono tentativi di rendere computazionalmente tractable questa correzione geometrica.

VI. L'entropia relativa nella teoria dei test statistici

Una lettura alternativa della divergenza KL, storicamente precedente all'interpretazione informazionale, emerge dalla teoria dei test di ipotesi. Se si vuole discriminare tra l'ipotesi nulla $H_0 : x \sim p$ e l'ipotesi alternativa $H_1 : x \sim q$ sulla base di n osservazioni indipendenti $x_1, \ldots, x$, il test del rapporto di verosimiglianza confronta:

$$\Lambda = \prod_{i=1}^{n} p(x_i)/q(x_i)$$

Per la legge dei grandi numeri, quando i dati sono generati da p:

$$(1/n) \log \Lambda \to D_{KL}(p \| q) \quad \text{quasi certamente}$$

mentre quando i dati sono generati da q la medesima media campionaria converge a $-D_{KL}(q \| p)$. La divergenza KL è dunque il tasso esponenziale al quale il rapporto di verosimiglianza cresce in modulo al crescere delle osservazioni.

Il lemma di Neyman-Pearson afferma che, tra tutti i test di livello α, il test del rapporto di verosimiglianza massimizza la potenza. Questa ottimalità universale del rapporto di verosimiglianza, connessa alla divergenza KL, suggerisce che l'informazione misuri qualcosa di obiettivo: la distinguibilità intrinseca di due distribuzioni, indipendentemente da qualsiasi scelta arbitraria del decisore.

VII. Entropia differenziale e le sue peculiarità

Nelle applicazioni moderne dell'intelligenza artificiale le variabili casuali di interesse sono quasi sempre continue. La generalizzazione dell'entropia a variabili continue richiede attenzione, perché presenta proprietà strutturalmente diverse da quelle del caso discreto.

Silvano Salvador

Per una variabile casuale continua X con densità p(x) rispetto alla misura di Lebesgue su $\mathbb{R}^n$, l'entropia differenziale è:

$$h(X) = -\int p(x) \log p(x)\, dx$$

A differenza dell'entropia discreta, h(X) può essere negativa — per una distribuzione concentrata su un intervallo molto piccolo — e non è invariante rispetto a trasformazioni invertibili della variabile. Se Y = f(X) con f differenziabile e invertibile:

$$h(Y) = h(X) + \mathbb{E}[\log |\det J_f(X)|]$$

dove J_f è lo jacobiano di f. Questa dipendenza dalla scala di misura è la manifestazione continua del fatto che la nozione di incertezza assoluta dipende dalla coordinatizzazione scelta. Per questa ragione, nelle applicazioni è spesso più fondamentale la divergenza KL tra distribuzioni continue:

$$D_{KL}(P \| Q) = \int p(x) \log [p(x) / q(x)]\, dx$$

che è invariante per trasformazioni invertibili di x — i termini aggiuntivi derivanti dall'entropia differenziale si cancellano nella differenza — ed è quindi una grandezza intrinseca alla coppia di distribuzioni.

Nell'ottimizzazione di modelli generativi neurali questa distinzione è cruciale: la funzione obiettivo per l'addestramento di un modello che vuole approssimare una distribuzione dati continua non può essere l'entropia differenziale della distribuzione del modello, ma deve essere la divergenza KL tra le due distribuzioni, o sue approssimazioni computazionalmente accessibili. Le differenze architetturali tra le principali famiglie di modelli generativi profondi — modelli variazionali, normalizing flows, modelli di diffusione, reti antagoniste generative — si riducono, in larga misura, a scelte diverse su quale approssimazione di questa divergenza ottimizzare, e con quali metodi.

Silvano Salvador

VIII. Il frame problem e l'incertezza quantificata

Prima di concludere questa prima parte è opportuno richiamare, in chiave informazionale, la critica più profonda al paradigma simbolico nei domini aperti: il problema del frame. McCarthy e Hayes avevano identificato già nel 1969 la difficoltà di specificare, in un sistema logico, tutto ciò che rimane invariato dopo un'azione. La circumscription di McCarthy e le varie estensioni della logica per gestire la non-monotonicità erano tentativi di risolvere questo problema all'interno del paradigma simbolico. Tutti questi tentativi condividevano una difficoltà strutturale: essi cercavano di codificare esplicitamente ciò che è vero per default usando regole simboliche — ma la nozione di normalità ha una struttura intrinsecamente probabilistica, non logica.

In un framework informazionale il problema del frame si trasforma: la distribuzione a priori sui possibili stati del mondo cattura, implicita ma precisamente, l'informazione su cosa è normale, cosa è probabile, cosa è eccezionale. Un evento che non modifica le proprietà rilevanti del mondo è un evento la cui occorrenza ha divergenza KL nulla rispetto alla distribuzione a priori sulle proprietà invarianti. Non è necessario enumerare esplicitamente cosa rimane uguale: la struttura probabilistica si occupa implicitamente di questa gestione.

Il prezzo da pagare è la difficoltà di costruire e aggiornare distribuzioni su spazi di alta dimensionalità — ma si tratta di una difficoltà computazionale, non di un'impossibilità logica. Nei domini dove la specifica esplicita del comportamento è realizzabile e la certificabilità è richiesta, il paradigma simbolico mantiene vantaggi strutturali. Nei domini dove il mondo è aperto, rumoroso e statisticamente strutturato, la rappresentazione probabilistica è l'unica architettura concettuale adeguata. Questa distinzione, mantenuta con rigore lungo tutto il saggio,

Silvano Salvador

è la premessa metodologica che rende possibile un confronto onesto tra i due paradigmi.

L'informazione come fondamento — Parte Seconda

La svolta bayesiana: inferenza, credenza e modelli generativi

I. Il teorema di Bayes come teoria della conoscenza

Vi è una tensione irrisolta al cuore della statistica frequentista che ne limita strutturalmente la capacità di modellare l'apprendimento. La tensione riguarda lo statuto ontologico dei parametri di un modello: nella concezione frequentista i parametri sono quantità fisse, sebbene sconosciute, e la probabilità si riferisce esclusivamente alla frequenza relativa di eventi in sequenze ipoteticamente infinite di ripetizioni. In questa concezione non ha senso attribuire una distribuzione di probabilità a un parametro: dire che θ ha probabilità 0.7 di essere maggiore di 5 è privo di significato, perché θ è un numero determinato, non un evento casuale.

Questa restrizione ontologica ha conseguenze metodologiche di grande portata. I metodi frequentisti possono stimare i parametri con stime puntuali come la massima verosimiglianza, o con intervalli di confidenza che si riferiscono alla variabilità degli stimatori sulla ripetizione ipotetica dell'esperimento, ma non possono descrivere lo stato di conoscenza del ricercatore riguardo ai parametri né aggiornarlo sistematicamente. Il

Silvano Salvador

concetto di probabilità che richiedono per descrivere l'incertezza sui parametri è precisamente quello che escludono per principio.

L'alternativa bayesiana, le cui radici risalgono al lavoro del reverendo Thomas Bayes pubblicato postumo nel 1763 e al trattato di Laplace del 1812, adotta una concezione radicalmente diversa: la probabilità è una misura del grado di credenza razionale, applicabile a qualsiasi proposizione incerta. Questa concezione soggettivista riceve una fondazione assiomatica dal teorema di Cox (1946): le uniche misure di credenza che soddisfano condizioni di coerenza — non contraddirsi nel combinare credenze indipendenti, essere consistenti con la logica proposizionale — sono le misure di probabilità che soddisfano gli assiomi di Kolmogorov. La probabilità bayesiana non è dunque una scelta filosofica arbitraria: è la struttura matematica necessaria per il ragionamento coerente sotto incertezza.

L'aggiornamento bayesiano è la regola normativa che descrive come lo stato di conoscenza debba essere modificato alla luce di nuove evidenze:

$$P(\theta \mid D) = P(D \mid \theta) \cdot P(\theta) / P(D)$$

dove $P(\theta)$ è la distribuzione a priori, $P(D \mid \theta)$ è la verosimiglianza, e $P(\theta \mid D)$ è la distribuzione a posteriori. La costante di normalizzazione $P(D) = \int P(D \mid \theta)\, P(\theta)\, d\theta$ è la verosimiglianza marginale o evidenza del modello. La formula è elementare; le sue implicazioni sono straordinarie. L'apprendimento, in questa visione, non è la ricerca di un parametro vero ma l'aggiornamento coerente di una distribuzione di credenza. Un sistema bayesiano mantiene rappresentata la propria incertezza residua, modulando le proprie previsioni in proporzione alla robustezza delle evidenze.

Silvano Salvador

II. La difficoltà dell'inferenza esatta e la derivazione rigorosa dell'ELBO

Il principale ostacolo computazionale all'applicazione del teorema di Bayes è il calcolo del denominatore P(D). Questa integrazione è in generale intrattabile non soltanto per ragioni di costo computazionale, ma per ragioni di struttura. Nei modelli di interesse pratico l'integrando $P(D \mid \theta) \cdot P(\theta)$ è una funzione altamente non convessa su uno spazio di dimensionalità che può superare il miliardo. Non esiste alcuna tecnica di quadratura applicabile in tale regime, e i metodi Monte Carlo via catene di Markov scalano esponenzialmente con la dimensione dello spazio quando il mixing delle catene diventa proibitivamente lento.

L'inferenza variazionale trasforma il problema: invece di calcolare $P(\theta \mid D)$, si cerca la distribuzione $q(\theta; \varphi)$ all'interno di una famiglia parametrica che minimizza la divergenza KL dalla vera posteriore:

$$\varphi^* = \operatorname{argmin}_\varphi D_{KL}(q(\theta; \varphi) \,\|\, P(\theta \mid D))$$

La chiave è spostare l'obiettivo su quantità calcolabili. Si parte dall'identità esatta, valida per qualsiasi q con lo stesso supporto di $P(\theta \mid D)$:

$$\log P(D) = \log \int P(D, \theta)\, d\theta$$

$$= \log \int q(\theta; \varphi) \cdot [P(D, \theta) / q(\theta; \varphi)]\, d\theta$$

$$= \log \mathbb{E}_{q(\theta;\varphi)}[P(D, \theta) / q(\theta; \varphi)]$$

Per la disuguaglianza di Jensen applicata alla funzione concava $\log(\cdot)$:

$$\log \mathbb{E}_q[P(D, \theta) / q(\theta; \varphi)] \geq \mathbb{E}_q[\log P(D, \theta) - \log q(\theta; \varphi)] =: \mathcal{L}(\varphi)$$

Silvano Salvador

La quantità $\mathcal{L}(\varphi)$ è l'Evidence Lower BOund (ELBO), un limite inferiore per log P(D) con uguaglianza se e soltanto se $q(\theta; \varphi) = P(\theta \mid D)$ quasi ovunque. La relazione esatta tra ELBO e log-evidenza è:

$$\log P(D) = \mathcal{L}(\varphi) + D_{KL}(q(\theta; \varphi) \parallel P(\theta \mid D))$$

poiché $D_{KL}(q \parallel P(\cdot|D)) = \mathbb{E}_q[\log q - \log P(\theta|D)] = \mathbb{E}_q[\log q - \log P(D,\theta) + \log P(D)] = -\mathcal{L}(\varphi) + \log P(D)$. Massimizzare $\mathcal{L}(\varphi)$ rispetto a φ è quindi equivalente a minimizzare $D_{KL}(q(\theta;\varphi) \parallel P(\theta|D))$, che era l'obiettivo originale, senza mai calcolare P(D).

Quando il modello ha forma fattorizzata $P(D, \theta) = P(D \mid \theta)\, P(\theta)$, l'ELBO si decompone in due termini interpretativamente distinti:

$$\mathcal{L}(\varphi) = \mathbb{E}_{q(\theta;\varphi)}[\log P(D \mid \theta)] - D_{KL}(q(\theta; \varphi) \parallel P(\theta))$$

Il primo termine è la verosimiglianza attesa — quanto bene il modello spiega i dati in media. Il secondo è la divergenza KL tra la distribuzione variazionale e il prior — il costo informazionale dell'aggiornamento delle credenze. Questa decomposizione accuratezza-complessità è la struttura dell'inferenza bayesiana resa computazionalmente trattabile: il sistema cerca la spiegazione delle osservazioni più accurata con il minor aggiornamento del prior, una formalizzazione del principio di parsimonia.

III. La direzione della divergenza KL: zero-forcing e mass-covering

Una questione che l'ottimizzazione variazionale impone di affrontare con cura riguarda la scelta della direzione della divergenza KL. L'inferenza variazionale standard minimizza $D_{KL}(q \parallel P(\cdot|D))$ — la divergenza forward — ma questa non è l'unica scelta possibile, e le conseguenze della scelta sono geometricamente precise.

Silvano Salvador

La divergenza $D_{KL}(q \| p)$ è zero-forcing: il suo gradiente rispetto a q tende a spingere q verso zero nei punti in cui p ha densità bassa ma non nulla, poiché il termine $q(\theta) \log[q(\theta)/p(\theta)] \to 0$ quando $q(\theta) \to 0$, non imponendo penalità per ignorare regioni di p. La distribuzione approssimante tende quindi a concentrarsi su una delle mode di p. Se p ha più mode separate — circostanza comune nelle distribuzioni posteriori di modelli complessi — la distribuzione variazionale si addensa tipicamente su una sola moda, perdendo l'informazione sulle altre.

La divergenza $D_{KL}(p \| q)$, al contrario, è mass-covering: se $p(\theta) > 0$ mentre $q(\theta) \to 0$, il termine $p(\theta) \log[p(\theta)/q(\theta)]$ diverge, penalizzando fortemente q per assegnare densità nulla a regioni a cui p assegna massa. La distribuzione approssimante tende a coprire tutto il supporto di p, a costo di assegnare densità a regioni di bassa probabilità, producendo un'approssimazione eccessivamente diffusa.

La letteratura ha proposto diverse alternative. Le α-divergenze di Rényi interpolano tra le due direzioni:

$$D_\alpha(P \| Q) = (1/(\alpha-1)) \log \int p(\theta)^\alpha \, q(\theta)^{(1-\alpha)} \, d\theta$$

con $D_\alpha \to D_{KL}(P \| Q)$ per $\alpha \to 1$ e $D_\alpha \to D_{KL}(Q \| P)$ per $\alpha \to 0$. Le importance-weighted autoencoders (IWAE) di Burda, Grosse e Salakhutdinov (2016) costruiscono un bound più stretto dell'ELBO standard usando campionamento per importanza con L campioni:

$$\log P(x) \geq \mathcal{L}_L(\varphi, \theta; x) = \mathbb{E}_{\{z_1,\ldots,z_L \sim q\}}[\log (1/L) \sum_{=1}^{L} P(x,z)/q(z|x;\varphi)]$$

Si verifica che $\mathcal{L}_L \geq \mathcal{L}_{L-1} \geq \ldots \geq \mathcal{L}_1 = \mathcal{L}_{ELBO}$ per qualsiasi L, con $\mathcal{L}_L \to \log P(x)$ per $L \to \infty$, avvicinando asintoticamente il bound alla log-evidenza esatta.

Silvano Salvador

IV. Il variational autoencoder: inferenza bayesiana profonda

Il variational autoencoder (VAE) di Kingma e Welling (2013) trasla la struttura dell'inferenza variazionale in un'architettura computazionale concreta per modelli con variabili latenti continue. Il modello generativo assume che ogni dato osservato $x \in \mathbb{R}^n$ sia generato secondo:

$z \sim P(z) = \mathcal{N}(z; 0, I_k)$

$x \mid z \sim P(x \mid z; \theta_dec)$

dove θ_dec parametrizza una rete neurale decodificatrice. La distribuzione a posteriori esatta $P(z \mid x; \theta_dec)$ richiede il calcolo di $P(x; \theta_dec) = \int P(x \mid z; \theta_dec)\, \mathcal{N}(z; 0, I_k)\, dz$ — l'evidenza intrattabile che si vuole evitare.

Il VAE introduce una distribuzione di riconoscimento $q(z \mid x; \varphi_enc)$ — l'encoder — e ottimizza l'ELBO:

$\mathcal{L}(\theta_dec, \varphi_enc; x) = \mathbb{E}_{q(z|x;\varphi_enc)}[\log P(x \mid z; \theta_dec)] - D_KL(q(z \mid x; \varphi_enc) \,\|\, P(z))$

Nella scelta standard l'encoder produce una gaussiana diagonale:

$q(z \mid x; \varphi_enc) = \mathcal{N}(z; \mu(x, \varphi_enc), \mathrm{diag}(\sigma^2(x, \varphi_enc)))$

Il termine di divergenza KL è calcolabile in forma chiusa. Indicando μ_i e σ_i^2 le componenti del vettore medio e della varianza diagonale, la derivazione procede calcolando esplicitamente:

$D_KL(q \,\|\, p) = \mathbb{E}_q[\log q(z)] - \mathbb{E}_q[\log p(z)]$

Per $q = \mathcal{N}(\mu, \mathrm{diag}(\sigma^2))$ e $p = \mathcal{N}(0, I)$:

$\mathbb{E}_q[\log q(z)] = -(k/2)\log(2\pi) - (1/2)\sum_i \log \sigma_i^2 - k/2$

$\mathbb{E}_q[\log p(z)] = -(k/2)\log(2\pi) - (1/2)\sum_i (\sigma_i^2 + \mu_i^2)$

Silvano Salvador

Sottraendo:

$$D_{KL}(\mathcal{N}(\mu, \mathrm{diag}(\sigma^2)) \,\|\, \mathcal{N}(0, I)) = \tfrac{1}{2} \sum_i (\sigma_i^2 + \mu_i^2 - 1 - \log \sigma_i^2)$$

Ogni addendo è non negativo per la disuguaglianza $\sigma^2 - 1 - \log \sigma^2 \geq 0$, con uguaglianza per $\sigma^2 = 1$, $\mu = 0$, fornendo una verifica geometrica della non-negatività della divergenza KL in questo caso.

Il termine di verosimiglianza attesa non ammette forma chiusa e richiede di differenziare attraverso un'operazione di campionamento — non differenziabile nel senso ordinario. Il trucco della riparametrizzazione risolve questo scrivendo:

$$z \sim \mathcal{N}(\mu, \mathrm{diag}(\sigma^2)) \iff z = \mu + \sigma \odot \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, I)$$

dove $\odot$ denota il prodotto di Hadamard. In questa forma z è una funzione differenziabile di μ e σ per qualsiasi ε fissato, e lo stimatore del gradiente diventa:

$$\nabla_{\varphi_enc} \mathbb{E}_q[\log P(x \mid z; \theta_dec)]$$

$$\approx \nabla_{\varphi_enc} \log P(x \mid \mu(x,\varphi_enc) + \sigma(x,\varphi_enc) \odot \hat{\varepsilon}; \theta_dec)$$

per un campione $\hat{\varepsilon} \sim \mathcal{N}(0, I)$. Questo è uno stimatore a varianza finita del gradiente, calcolabile con retropropagazione ordinaria. La condizione sufficiente per l'applicabilità del trucco è l'esistenza di una variabile ε con distribuzione fissa e di una funzione differenziabile g tale che $z = g(\varepsilon, \varphi)$ — condizione soddisfatta oltre che dalla gaussiana dalla distribuzione di Gumbel, dalla famiglia esponenziale con trasformazioni di inversione della CDF, e da qualsiasi distribuzione parametrizzabile come trasformazione differenziabile di rumore semplice.

V. Il collasso della distribuzione a posteriori: analisi formale

Silvano Salvador

Il posterior collapse è un fenomeno strutturale che emerge nell'addestramento dei VAE con decoder sufficientemente espressivi. Formalmente, per alcune dimensioni j del vettore latente, la distribuzione dell'encoder $q(z \mid x; \varphi_enc)$ converge alla distribuzione a priori $P(z) = \mathcal{N}(0,1)$ indipendentemente dall'input x.

La meccanica del collasso si analizza considerando la decomposizione dell'ELBO per ciascuna dimensione latente separatamente. Se il decoder apprende a ricostruire x ignorando la j-esima dimensione latente — $\partial \log P(x|z)/\partial z \approx 0$ — allora il termine di verosimiglianza attesa è approssimativamente costante rispetto ai parametri dell'encoder per quella dimensione. L'ottimizzazione dell'ELBO per la dimensione j si riduce alla minimizzazione di $D_KL(q(z|x) \| P(z))$, il cui minimo è zero e viene raggiunto quando $q(z|x) = \mathcal{N}(0,1)$ per ogni x.

Diverse soluzioni architetturali sono state proposte. Il KL annealing pesa il termine di penalità con un coefficiente $\beta < 1$ inizialmente, portandolo progressivamente a 1. Il β-VAE di Higgins et al. (2017) fissa $\beta > 1$ per incoraggiare rappresentazioni disentangled. Le varianti con free bits di Kingma et al. (2016) impongono un contributo minimo garantito per ciascuna dimensione latente:

$$\mathcal{L}_free = \mathbb{E}_q[\log P(x \mid z)] - \sum \max(\lambda, D_KL(q(z|x) \| P(z)))$$

dove $\lambda > 0$ è una soglia che impedisce il collasso escludendo dalla penalizzazione le dimensioni il cui contributo KL supera già la soglia.

VI. La critica del paradigma simbolico attraverso la lente variazionale

L'opposizione tra inferenza variazionale e sistemi simbolici non è soltanto tecnica; è concettuale. Un sistema esperto simbolico — si pensi a MYCIN, sviluppato a Stanford negli anni settanta per la diagnosi delle

Silvano Salvador

malattie infettive — codificava la conoscenza medica in regole del tipo SE … ALLORA … con coefficienti numerici che i progettisti chiamavano fattori di certezza. Questi coefficienti non erano probabilità nel senso formale del termine: erano parametri euristici che catturavano il giudizio clinico nella misura in cui esso poteva essere esplicitato verbalmente.

È importante notare che i progettisti di MYCIN erano consapevoli di questa limitazione e la discutevano esplicitamente. Il problema fondamentale di MYCIN non era l'eccessiva semplicità delle regole, ma la struttura epistemologica che esso presupponeva: la conoscenza come insieme finito e articolabile di relazioni tra concetti discreti. Questa assunzione era, alla luce di ciò che sappiamo sull'apprendimento biologico, radicalmente inadeguata per i domini aperti. La conoscenza che un medico esperto usa nell'atto diagnostico è distribuita, tacita, contestuale — è il tacit knowledge di cui aveva parlato Michael Polanyi nel 1966.

L'inferenza variazionale profonda offre un modello computazionale del tacit knowledge: una distribuzione di probabilità su configurazioni latenti, appresa dall'esperienza, che cattura strutture regolari nei dati senza che esse vengano mai esplicitate in forma proposizionale. Quando un modello bayesiano profondo riconosce strutture in un input, esso esegue un'inferenza sulla distribuzione latente, ponderando implicitamente l'intera struttura correlazionale appresa dall'addestramento.

Va però segnalata una limitazione importante che il confronto con MYCIN tende a oscurare. Il tacit knowledge di Polanyi non è semplicemente un pattern statistico in dati di comportamento: è una forma di sapere incorporato — embodied knowledge — che emerge

Silvano Salvador

dall'interazione sensomotoria prolungata con un dominio fisico. Un modello addestrato su descrizioni testuali di casi clinici non ha accesso alla dimensione sensomotoria del sapere clinico, e la qualità delle sue prestazioni nei casi atipici dipende criticamente da quanto la distribuzione dei casi nel corpus di addestramento copra il territorio rilevante. La distribuzione statistica delle manifestazioni di una competenza non è la competenza stessa: questa distinzione è epistemologicamente rilevante e non si lascia dissolvere con l'argomento distributivista.

VII. L'inferenza in-context come inferenza bayesiana implicita

Uno degli aspetti più intriganti dei grandi modelli linguistici contemporanei è la capacità di apprendimento in-context: adattare il proprio comportamento sulla base di esempi forniti nel contesto di ingresso, senza alcun aggiornamento dei parametri. Un modello vede tre esempi di una mappatura input-output mai incontrata durante l'addestramento, e produce un output corretto per un quarto input.

Un'interpretazione bayesiana rigorosa di questo fenomeno è stata proposta da Xie e colleghi (2021). Essi hanno mostrato che, se il corpus di addestramento è generato da un processo latente in cui ogni documento è prodotto da un concetto latente c campionato da P(c), e se il modello apprende la distribuzione congiunta sui documenti, allora l'inferenza in-context realizza implicitamente l'inferenza bayesiana sulla distribuzione a posteriori del concetto latente:

$$P(c \mid x_1, y_1, \ldots, x , y) \propto P(c) \cdot \prod_i P(x_i, y_i \mid c)$$

La previsione del prossimo token è dunque una marginalizzazione sul concetto latente:

Silvano Salvador

$P(y_next \mid x_1, y_1, \ldots, x , x_next) = \int P(y_next \mid x_next, c) \cdot P(c \mid x_1, y_1, \ldots, x) \, dc$

Il modello linguistico realizza implicitamente questo calcolo bayesiano senza alcuna struttura simbolica esplicita che implementi l'aggiornamento della credenza. La logica bayesiana emerge come comportamento della distribuzione appresa, non come meccanismo primitivo.

VIII. La questione del prior e la regolarizzazione implicita

L'inferenza bayesiana rigorosa richiede la specifica di una distribuzione a priori $P(\theta)$ prima di osservare i dati. Nell'apprendimento profondo la distribuzione a priori sui parametri è spesso implicita piuttosto che esplicita. La regolarizzazione L_2 — la penalità proporzionale alla norma quadratica dei pesi — corrisponde esattamente, nella prospettiva bayesiana, all'uso di un prior gaussiano isotropo sui parametri: il massimizzatore della verosimiglianza penalizzata è il punto di massima densità a posteriori (MAP) sotto tale prior. La regolarizzazione L_1 corrisponde a un prior laplaciano. L'interpretazione bayesiana di queste tecniche, comunemente usate senza esplicitarne la natura probabilistica, illumina il fatto che non esiste apprendimento privo di assunzioni a priori: ogni procedura di ottimizzazione impone implicitamente una struttura a priori.

La scelta del prior per le variabili latenti nel VAE — la gaussiana standard $\mathcal{N}(0, I)$ — non è arbitraria. Essa è il prior massima entropia soggetto al vincolo che la media sia nulla e che la varianza per ciascuna dimensione sia unitaria, come segue dal principio di Jaynes. È dunque la scelta meno presupposta compatibile con le proprietà di normalizzazione e isotropia.

Silvano Salvador

IX. L'incompletezza bayesiana e il problema della misspecification

La misspecification è il limite strutturale più profondo del framework bayesiano, e merita un'analisi formale precisa. Si assuma che la vera distribuzione generativa dei dati sia P^* — ignota e non rappresentabile nella classe di modelli adottata. L'analisi asintotica della distribuzione a posteriori sotto misspecification è governata dal teorema di Bernstein-von Mises generalizzato nella versione di Kleijn e van der Vaart (2012).

Il risultato centrale è che la distribuzione a posteriori converge, sotto P^*, attorno al punto di pseudo-verità:

$$\theta_0 = \text{argmin}_\theta \, D_{KL}(P^* \,\|\, P(\cdot \mid \theta))$$

ovvero il parametro che minimizza la divergenza KL dalla distribuzione vera alla classe di modelli considerata. La distribuzione a posteriori non converge a una gaussiana centrata sul vero θ — che non è nemmeno definito quando P^* non appartiene alla classe — bensì a una gaussiana centrata su θ_0 con matrice di covarianza sandwich:

$$\Sigma_{post} \approx (1/n)\, G(\theta_0)^{-1}\, J(\theta_0)\, G(\theta_0)^{-1}$$

dove $G(\theta_0)$ è la matrice di Fisher del modello e

$$J(\theta_0) = \mathbb{E}_{P^*}[\nabla \log P(x \mid \theta_0) \; \nabla \log P(x \mid \theta_0)^T]$$

è la matrice di informazione osservata sotto la vera distribuzione. Nei modelli correttamente specificati $G = J$ e la formula si riduce all'espressione classica. In caso contrario le due matrici differiscono, e ignorare questa discrepanza produce intervalli di credibilità con copertura frequentista non nominale.

Silvano Salvador

La struttura di questo risultato chiarisce che il valore epistemologico dell'inferenza bayesiana è condizionato alla qualità del modello: un sistema che usa il modello sbagliato produce incertezza calibrata rispetto al modello, non rispetto alla realtà. La misspecification non è risolvibile dall'interno del framework bayesiano: richiede strumenti esterni come la validazione predittiva leave-one-out, il WAIC, o le procedure di inferenza semiparametrica.

Vale la pena sottolineare che la misspecification è un problema irriducibile per qualsiasi sistema computazionalmente trattabile. Per ogni classe di modelli finita, esiste sempre una distribuzione generativa reale che non è catturata da quella classe — questo è il contenuto del no-free-lunch theorem nella stima di densità. Espandere la classe sposta il problema: un modello abbastanza espressivo da approssimare qualsiasi distribuzione computabile è il predittore di Solomonoff, non calcolabile. Il paradigma informazionale offre strumenti per quantificare e gestire la misspecification; non offre strumenti per eliminarla.

X. RLHF come inferenza bayesiana applicata: la connessione con il Saggio III

Il reinforcement learning da feedback umano, che il prossimo saggio analizzerà in dettaglio come metodo di allineamento, è strutturalmente un problema di inferenza bayesiana: trovare la distribuzione a posteriori $P(\theta \mid D)$ sui valori umani θ dati i dati di comparazione D. In questa sezione si mostra come gli strumenti della Parte Seconda si applichino direttamente a questo problema, stabilendo il ponte informazionale tra i due saggi.

Il modello generativo è il seguente. I valori latenti dell'utente $\theta \in \Theta$ generano le comparazioni osservate: dato un prompt x e due risposte

Silvano Salvador

y_w, y_l, la probabilità che y_w sia preferita a y_l è modellata secondo Bradley-Terry:

$$P(y_w \succ y_l \mid x; \theta) = \sigma(R_\theta(x, y_w) - R_\theta(x, y_l))$$

dove R_θ è la funzione di utilità parametrizzata da θ. La distribuzione a posteriori su θ dati N confronti è:

$$P(\theta \mid D) \propto \prod_{i=1}^{N} P(y_w^{(i)} \succ y_l^{(i)} \mid x^{(i)}; \theta) \cdot P(\theta)$$

Il reward model R_φ del RLHF è un'approssimazione variazionale di questa a posteriori. Anziché mantenere la distribuzione completa $P(\theta \mid D)$ — computazionalmente intrattabile per spazi parametrici di miliardi di dimensioni — si apprende una stima puntuale $\hat{\theta} = \text{argmax } P(\theta \mid D)$ attraverso la massimizzazione della log-verosimiglianza penalizzata. Questa è esattamente la MAP estimation discussa nel paragrafo I di questa Parte, con il prior $P(\theta)$ che corrisponde alla regolarizzazione implicita dell'architettura neurale.

Il limite strutturale di questa approssimazione puntuale è precisamente il primo tipo di misspecification bayesiana discussa nel paragrafo IX: la stima MAP collassa la distribuzione a posteriori su un singolo punto, perdendo l'incertezza sulle preferenze. In presenza di valutatori in disaccordo — che è la norma, non l'eccezione, nei dataset di preferenze reali — la distribuzione a posteriori $P(\theta \mid D)$ è multimodale: esistono più sistemi di valori ugualmente compatibili con le osservazioni. Il reward model scalare apprende la media di questa distribuzione, che può non corrispondere a nessun sistema di valori reale. L'approccio bayesiano corretto — mantenere un ensemble di reward model $R_{\theta 1}, ..., R_{\theta K}$ che campioni la distribuzione a posteriori — produce agenti più robusti e correggibili, a costo di complessità computazionale significativamente maggiore.

Silvano Salvador

Questa connessione ha un'implicazione immediata per il tema dell'incertezza epistemica discusso nel paragrafo VIII di questa Parte: un agente allineato dovrebbe propagare l'incertezza sulle preferenze umane, non colassarla. La calibrazione dell'incertezza nel RLHF — sapere quando il sistema non sa quali valori applicare perché la distribuzione a posteriori è ampia — è un requisito fondamentale di sicurezza che i metodi attuali non soddisfano sistematicamente, e che il programma di ricerca bayesiano aperto identificato nel paragrafo IX si propone di affrontare.

XI. Distribuzioni a priori gerarchiche e meta-apprendimento bayesiano

Un'estensione naturale del framework variazionale è l'introduzione di distribuzioni a priori gerarchiche, in cui i parametri del prior sono a loro volta distribuiti secondo un iperpriore. La struttura gerarchica permette al modello di adattare il prior sulla base dei dati, catturando la struttura statistica comune a famiglie di problemi correlati.

Nel meta-apprendimento bayesiano, il compito è apprendere una distribuzione a priori $P(\theta)$ tale che, dato un piccolo insieme di esempi da un nuovo compito, l'inferenza bayesiana produca rapidamente una buona distribuzione a posteriori. Il framework MAML di Finn, Abbeel e Levine (2017) può essere interpretato in questa chiave: l'inizializzazione dei parametri appresa da MAML è un prior implicito tale che pochi passi di gradiente dalla condizione iniziale producano buone soluzioni per nuovi compiti della stessa famiglia. L'interpretazione bayesiana del meta-apprendimento non è retorica: essa collega la struttura formale dell'inferenza bayesiana alle dinamiche di ottimizzazione delle reti neurali, mostrando come le proprietà statistiche dell'apprendimento emergano dai meccanismi computazionali dell'ottimizzazione.

Silvano Salvador

L'informazione come fondamento — Parte Terza

Termodinamica dell'informazione, score matching e la genealogia dei modelli di diffusione

I. La convergenza tra fisica statistica e teoria dell'informazione

Il XIX secolo ha prodotto, in campi apparentemente lontani, due teorie dell'incertezza strutturalmente isomorfe: la meccanica statistica di Boltzmann e la teoria dell'informazione che Shannon avrebbe formalizzato un secolo dopo. L'isomorfismo non è accidentale né meramente formale. Esso riflette una comunanza profonda nel problema affrontato: come distribuire probabilità su un insieme di configurazioni possibili in presenza di vincoli parziali sull'informazione disponibile.

Prima di percorrere questa convergenza occorre però chiarire una distinzione epistemologica che la monografia originalità trascura e che ha conseguenze di rilievo per l'interpretazione dei modelli generativi: la distinzione tra l'entropia come grandezza fisica e l'entropia come grandezza epistemica. La tradizione boltzmanniana identifica l'entropia termodinamica di un sistema con il numero di microstati compatibili con il suo macrostato — una quantità che dipende dallo stato oggettivo del sistema. La tradizione shannoniana — e in modo ancora più esplicito quella di Jaynes — identifica l'entropia con l'incertezza soggettiva di un agente rispetto al sistema. Le due concezioni producono previsioni numericamente coincidenti in molti contesti fisici, il che ha alimentato una confusione che rischia di trasportarsi, non elaborata, nell'interpretazione dei modelli di apprendimento.

Silvano Salvador

La distinzione non è puramente filosofica. Nel contesto dei modelli generativi, l'entropia che si massimizza nella costruzione dei prior è sempre epistemica nel senso di Jaynes: è la misura dell'incertezza del modello rispetto alla distribuzione dei dati, non una proprietà intrinseca di quei dati. L'energia libera variazionale di Friston, discussa nella sezione conclusiva di questa parte, è parimenti epistemica: essa misura l'incertezza di un agente cognitivo, non uno stato termodinamico del cervello. Mantenere questa distinzione evita equivoci sull'interpretazione dei risultati.

II. La distribuzione di Boltzmann come soluzione del problema variazionale

Consideriamo un sistema fisico con stati discreti $x \in \mathcal{X}$ e funzione di energia $E : \mathcal{X} \to \mathbb{R}$. La termodinamica statistica di Gibbs afferma che, all'equilibrio termico a temperatura T, la distribuzione degli stati è la distribuzione di Boltzmann-Gibbs:

$$p^*(x) = \exp(-E(x)/kT) \,/\, Z$$

dove k è la costante di Boltzmann e $Z = \sum_x \exp(-E(x)/kT)$ è la funzione di partizione. Perché questa distribuzione? Il principio di massima entropia di Jaynes fornisce la risposta: tra tutte le distribuzioni di probabilità compatibili con il vincolo che il valore atteso dell'energia sia uguale a un valore osservato μ_E, la distribuzione di Boltzmann è quella che massimizza l'entropia di Shannon.

La derivazione procede attraverso il metodo dei moltiplicatori di Lagrange. Il Lagrangiano da massimizzare è:

$$\mathcal{J}(p, \beta, \lambda_0) = -\sum_x p(x) \log p(x) - \beta(\sum_x p(x)\, E(x) - \mu_E) - \lambda_0(\sum_x p(x) - 1)$$

La derivata funzionale rispetto a p(x) è:

Silvano Salvador

$$-\log p(x) - 1 - \beta E(x) - \lambda_0 = 0$$

da cui $p^*(x) = \exp(-1 - \lambda_0) \cdot \exp(-\beta E(x))$. La costante di normalizzazione fissa $\exp(1 + \lambda_0) = Z(\beta) = \sum_x \exp(-\beta E(x))$, producendo:

$$p^*(x) = \exp(-\beta E(x)) / Z(\beta)$$

con $\beta = 1/(kT)$ determinato dal vincolo sull'energia media. La distribuzione di Boltzmann non è dunque un postulato fisico: è la conseguenza del principio di parsimonia informazionale applicato al vincolo sull'energia media.

La funzione di partizione $Z(\beta)$ svolge un ruolo generativo di straordinaria centralità nella fisica statistica e nell'apprendimento automatico. Le grandezze termodinamiche si ottengono come derivate logaritmiche di Z rispetto ai parametri:

$$\langle E \rangle = -\partial \log Z / \partial \beta$$

$$\mathrm{Var}[E] = \partial^2 \log Z / \partial \beta^2$$

$$F = -kT \log Z \quad \text{(energia libera di Helmholtz)}$$

L'energia libera $F = \langle E \rangle - TS$ — dove $S = -k \sum_x p^*(x) \log p^*(x)$ è l'entropia — è la grandezza termodinamica che un sistema all'equilibrio minimizza a temperatura fissa. Questa minimizzazione è esattamente il duale della massimizzazione dell'entropia sotto il vincolo sull'energia media: i due principi — massima entropia e minima energia libera — sono matematicamente equivalenti.

III. La diade demone di Maxwell e principio di Landauer

Il demone di Maxwell, proposto nel 1867, è un esperimento mentale che sembrava contraddire il secondo principio della termodinamica.

Silvano Salvador

Consideriamo un gas diviso in due camere comunicanti attraverso una piccola apertura. Un agente microscopico — il demone — osserva le molecole che si avvicinano all'apertura e apre o chiude un battente in modo da far passare soltanto le molecole veloci dalla camera A alla camera B, e soltanto quelle lente dalla B alla A. Dopo un tempo sufficiente la camera B sarà più calda della A, e il demone avrebbe apparentemente ridotto l'entropia del sistema senza compiere lavoro.

La risoluzione del paradosso avanzata da Rolf Landauer nel 1961 e rifinita da Charles Bennett nel 1982 è la seguente: il demone deve misurare e memorizzare lo stato di ogni molecola prima di decidere se aprire o chiudere il battente. Questo processo di acquisizione di informazione è reversibile e non richiede dissipazione di energia. Tuttavia, dopo aver elaborato un numero sufficiente di molecole, la memoria del demone si esaurisce. Per continuare a operare il demone deve cancellare i vecchi record dalla sua memoria. Ed è qui che il secondo principio interviene: cancellare un bit di informazione da una memoria fisica richiede la dissipazione di almeno kT log 2 joule di energia sotto forma di calore nell'ambiente. Questo è il principio di Landauer.

Il principio di Landauer stabilisce un legame quantitativo tra informazione e termodinamica: la cancellazione irreversibile di un bit di informazione ha un costo energetico minimo pari a kT log 2. Occorre però precisare la natura di questo risultato per evitarne un'interpretazione troppo forte. Il limite di Landauer è un risultato termodinamico, non un teorema di teoria dell'informazione: esso riguarda operazioni fisiche su supporti fisici, non operazioni astratte su stringhe di simboli. Come ha argomentato Bennett (2003), il processo di misurazione è logicamente reversibile e non ha costo energetico di principio; è la cancellazione —

Silvano Salvador

un'operazione logicamente irreversibile, che mappa molti stati iniziali in uno stato finale — a richiedere dissipazione.

La distinzione tra reversibilità logica e reversibilità termodinamica ha importanza diretta per i modelli generativi. I processi di diffusione stocastica che si discuteranno nelle sezioni seguenti sono logicamente irreversibili nel senso che il processo forward perde informazione sull'input iniziale. Ma la teoria di Anderson (1982) mostra che esiste un processo reverse che ha le stesse distribuzioni marginali, suggerendo che l'irreversibilità informazionale del processo forward non sia necessariamente irreversibilità termodinamica: la struttura matematica del processo reverse recupera l'informazione perduta attraverso la score function, non attraverso una reversione fisica del processo fisico.

IV. Le reti di Hopfield e la memoria come minimizzazione dell'energia

Le reti di Hopfield, proposte da John Hopfield nel 1982, sono il primo esempio di sistema neurale in cui la funzione di energia della fisica statistica diventa l'oggetto di ottimizzazione esplicito. Una rete di Hopfield è un sistema di n neuroni binari $\sigma_i \in \{-1, +1\}$ con connessioni simmetriche $W_{ij} = W_{ji}$ e senza auto-connessioni $W_{ii} = 0$. La funzione di energia è:

$$E(\sigma) = -\tfrac{1}{2} \sum_{i} \sum_{j} W_{ij} \sigma_i \sigma_j - \sum_i b_i \sigma_i$$

L'aggiornamento asincrono di un neurone i secondo la regola:

$$\sigma_i \leftarrow \mathrm{sgn}\left(\sum_j W_{ij} \sigma_j + b_i\right)$$

riduce monotonicamente $E(\sigma)$ o la lascia invariata, garantendo la convergenza a un minimo locale. La memoria della rete sono i pattern memorizzati come minimi locali della funzione di energia: quando si

Silvano Salvador

presenta alla rete un pattern parziale o corrotto, la dinamica di aggiornamento porta la rete verso il minimo più vicino, completando o correggendo il pattern originale.

La capacità di memorizzazione delle reti di Hopfield classiche — il numero di pattern che possono essere memorizzati prima che il recupero diventi inaffidabile — è limitata a circa 0.138n per n neuroni, come dimostrato da Amit, Gutfreund e Sompolinsky (1985) attraverso l'analisi della meccanica statistica della rete. Questo limite emerge dalla struttura delle interazioni a coppie che definisce la funzione di energia.

Le Modern Hopfield Networks di Ramsauer e colleghi (2020) riformulano il problema con una funzione di energia esponenzialmente espressiva. Per n pattern memorizzati $\xi^1, \ldots, \xi^n \in \mathbb{R}^d$, la funzione di energia per uno stato di query $x \in \mathbb{R}^d$ è:

$$E(x) = -\text{lse}(\beta, \Xi^T x) + \tfrac{1}{2} \|x\|^2 + \beta^{-1} \log n$$

dove $\text{lse}(\beta, z) = \beta^{-1} \log \sum_i \exp(\beta z_i)$ è il log-sum-exp con parametro di inversa temperatura β, e $\Xi = (\xi^1, \ldots, \xi^n)$ è la matrice dei pattern. La regola di aggiornamento per x è:

$$x_new = \Xi \, \text{softmax}(\beta \, \Xi^T x)$$

che converge globalmente con alta probabilità a un punto fisso vicino al pattern più simile a x. La capacità di memorizzazione scala esponenzialmente con la dimensione dei pattern — $O(\exp(d/2))$ — eliminando il collo di bottiglia delle reti classiche.

Il legame con il meccanismo di attenzione dei transformer è diretto. L'aggiornamento delle Hopfield Networks moderne è formalmente identico all'attenzione scaled dot-product di Vaswani e colleghi (2017) quando si identificano i valori V come pattern da recuperare, le chiavi K come pattern su cui misurare la similarità con la query Q, e $\beta = 1/\sqrt{d_k}$

Silvano Salvador

come parametro di temperatura. Ogni strato di attenzione è quindi un singolo passo di aggiornamento di una Hopfield Network moderna — un'operazione di recupero da memoria per contenuto — interpretazione che unifica architettura e teoria della memoria in un'unica struttura.

V. Score matching: fondamenta e derivazione rigorosa

La stima della score function — il gradiente della log-densità $\nabla_x \log p(x)$ — è il meccanismo centrale dei modelli di diffusione generativa. La score function di una distribuzione p è la funzione $s(x) : \mathbb{R}^n \to \mathbb{R}^n$ definita da $s(x) = \nabla_x \log p(x)$. Se p ha la forma di Gibbs $p(x) \propto \exp(-E(x)/kT)$, la score function è $s(x) = -\nabla_x E(x)/kT$ — proporzionale alla forza che il potenziale esercita sulla particella — e non richiede la conoscenza della funzione di partizione Z.

La dinamica di Langevin, utilizza la score function per campionare da p senza conoscerne la normalizzazione:

$$x_{t+1} = x_t + \varepsilon \nabla_x \log p(x_t) + \sqrt{(2\varepsilon)}\, z_t, \quad z_t \sim \mathcal{N}(0, I)$$

Sotto condizioni di regolarità, la distribuzione stazionaria di questa catena di Markov è p, indipendentemente dal punto di partenza x_0. La score function è il vettore di deriva che spinge le traiettorie verso le regioni di alta densità, mentre il termine di rumore $\sqrt{(2\varepsilon)}\, z_t$ garantisce l'esplorazione del supporto.

Il problema della stima della score function da dati è che la formula $s(x) = \nabla_x \log p(x)$ non è accessibile direttamente: richiederebbe conoscere p, che è precisamente ciò che si vuole imparare. Hyvärinen (2005) propose l'obiettivo di score matching:

$$J(\theta) = \tfrac{1}{2}\, \mathbb{E}_{p_{data}(x)}[\|s_\theta(x) - \nabla_x \log p_{data}(x)\|^2]$$

Silvano Salvador

Per la disuguaglianza di Fisher, la divergenza tra la score function del modello e quella dei dati è:

$$J(P_{data} \,\|\, P_{model}) = \mathbb{E}_{P_{data}}[\|\nabla_x \log p_{data}(x) - s_\theta(x)\|^2]$$

Integrando per parti rispetto alle coordinate, Hyvärinen dimostrò che, a meno di costanti indipendenti da θ:

$$J(\theta) \equiv \mathbb{E}_{p_{data}(x)}[\mathrm{tr}(\nabla_x s_\theta(x)) + \tfrac{1}{2}\|s_\theta(x)\|^2]$$

L'obiettivo implicito non richiede la conoscenza di p_{data}: dipende soltanto da campioni di p_{data} e dalla traccia dello jacobiano di s_θ. La traccia può essere stimata stocasticamente con l'estimatore di Hutchinson: $\mathrm{tr}(J) \approx v^T J v$ per un vettore casuale $v \sim \mathcal{N}(0, I)$, riducendo il costo computazionale da $O(n^2)$ a $O(n)$ per valutazione.

Vincent (2011) dimostrò che il denoising score matching — addestrare s_θ a recuperare il segnale originale da versioni rumorose — è equivalente allo score matching implicito sulla distribuzione smoothed $p_\sigma(\tilde{x}) = \int p_{data}(x)\, \mathcal{N}(\tilde{x}; x, \sigma^2 I)\, dx$. La connessione centrale è che la score function della distribuzione smoothed ha la forma analitica:

$$\nabla_{\tilde{x}} \log p_\sigma(\tilde{x}) = -(\tilde{x} - \mathbb{E}[x \mid \tilde{x}]) / \sigma^2$$

La derivazione rigorosa utilizza l'identità $p_{data}(x)\, q_\sigma(\tilde{x} \mid x) = p_\sigma(\tilde{x})\, p(x \mid \tilde{x})$ e il teorema di Bayes per la densità condizionale:

$$\nabla_{\tilde{x}} \log p_\sigma(\tilde{x}) = \nabla_{\tilde{x}} \log \int p_{data}(x)\, \mathcal{N}(\tilde{x}; x, \sigma^2 I)\, dx$$

$$= \left[\int p_{data}(x)\, \nabla_{\tilde{x}}\, \mathcal{N}(\tilde{x}; x, \sigma^2 I)\, dx\right] / p_\sigma(\tilde{x})$$

$$= \left[\int p_{data}(x) \cdot (-(\tilde{x} - x)/\sigma^2) \cdot \mathcal{N}(\tilde{x}; x, \sigma^2 I)\, dx\right] / p_\sigma(\tilde{x})$$

$$= -(\tilde{x} - \mathbb{E}[x \mid \tilde{x}]) / \sigma^2$$

Silvano Salvador

Addestrare $s_\theta(\tilde{x})$ a predire il segnale denoised $\mathbb{E}[x \mid \tilde{x}]$ equivale dunque a stimare la score function di p_σ scalata per $-\sigma^2$. Questa equivalenza trasforma il problema di stima della densità — intrattabile in alta dimensione — nel problema di denoising — affrontabile con regressione su dati rumorosi.

VI. Dalle equazioni differenziali stocastiche ai modelli di diffusione

L'unificazione matematica dei modelli di diffusione generativa attraverso la teoria delle equazioni differenziali stocastiche, proposta da Song e colleghi (2021), rivela la struttura profonda che connette processi fisici di diffusione e apprendimento generativo. Si consideri una SDE di Itô in forma generale:

$$dX_t = f(X_t, t)\,dt + g(t)\,dW_t$$

dove $f(X_t, t)$ è il termine di drift, $g(t)$ è il coefficiente di diffusione scalare, e W_t è un moto browniano standard n-dimensionale. L'evoluzione della densità di probabilità $p_t(x)$ del processo è governata dall'equazione di Fokker-Planck:

$$\partial_t p_t = -\mathrm{div}(f(x,t)\, p_t) + \tfrac{1}{2}\, g(t)^2\, \Delta\, p_t$$

Per il processo forward di riferimento dei DDPM, $f(x,t) = -\beta(t)/2 \cdot x$ e $g(t) = \sqrt{\beta(t)}$, la soluzione del marginal è gaussiana con:

$$q(x_t \mid x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}\, x_0, (1-\bar{\alpha}_t)\, I)$$

dove $\bar{\alpha}_t = \exp(-\int_0^t \beta(s)\, ds)$. Questa formula si ricava direttamente dalla soluzione della ODE per la media e la varianza del processo di Ornstein-Uhlenbeck, come mostrato in dettaglio nell'Appendice II.

Silvano Salvador

Il risultato fondamentale di Anderson (1982) afferma che a ogni processo di diffusione forward corrisponde un processo reverse con le stesse distribuzioni marginali:

$$dX_t = [-f(X_t, T-t) + g(T-t)^2 \nabla_x \log p_{T-t}(X_t)]\, dt + g(T-t)\, dW_t$$

Il termine aggiuntivo $g(t)^2 \nabla_x \log p_t(x)$ — la score function della distribuzione marginale al tempo t scalata per $g(t)^2$ — trasforma il drift del processo reverse. L'intuizione è precisa: il processo reverse deve rimuovere il rumore che il processo forward ha aggiunto, e la score function indica la direzione verso le regioni di alta densità della distribuzione corrente, ovvero la direzione del segnale al di sotto del rumore.

Sostituendo la stima neurale $s_\theta(x, t) \approx \nabla_x \log p_t(x)$ nella SDE reverse, l'obiettivo di addestramento dei DDPM — la perdita L_{simple} di Ho e colleghi (2020) — diventa:

$$L_{simple}(\theta) = \mathbb{E}_{t, x_0, \varepsilon}[\|\varepsilon - \varepsilon_\theta(\sqrt{\bar{\alpha}_t}\, x_0 + \sqrt{(1-\bar{\alpha}_t)}\, \varepsilon, t)\|^2]$$

dove ε_θ predice il rumore aggiunto, che è proporzionale alla score function negata: $\varepsilon_\theta(x_t, t) \approx -\sqrt{(1-\bar{\alpha}_t)}\, s_\theta(x_t, t)$. La derivazione di questa equivalenza si trova nell'Appendice II.

La Probability Flow ODE — la versione deterministica del processo reverse — è:

$$dx/dt = f(x, T-t) - \tfrac{1}{2}\, g(T-t)^2 \nabla_x \log p_{T-t}(x)$$

$$= -\beta(T-t)/2 \cdot (x + s_\theta(x, T-t))$$

Questa ODE può essere integrata con solvers numerici adattativi di ordine elevato (Dormand-Prince, Adams-Moulton), riducendo drasticamente il numero di valutazioni della rete necessarie per la

Silvano Salvador

generazione accurata rispetto alla simulazione stocastica della SDE. Questa è la base teorica dei metodi DDIM e dei successivi solvers ODE per modelli di diffusione.

VII. Il principio di energia libera variazionale di Friston: struttura e limiti

Karl Friston ha proposto il principio di energia libera variazionale come teoria unificante della cognizione adattiva. La struttura matematica è la seguente. Per un sistema cognitivo che riceve input sensoriali $o \in \mathbb{R}^m$ e mantiene una rappresentazione delle cause latenti $\vartheta \in \mathbb{R}^d$ sotto forma di distribuzione variazionale $q(\vartheta; \varphi)$, l'energia libera variazionale è:

$$F = D_{KL}(q(\vartheta; \varphi) \,\|\, P(\vartheta \mid o)) - \log P(o)$$

$$= \mathbb{E}_q[\log q(\vartheta; \varphi) - \log P(o, \vartheta)]$$

Poiché $D_{KL} \geq 0$, si ha $F \geq -\log P(o)$, ovvero F è un upper bound sulla sorpresa $-\log P(o)$. La scomposizione:

$$F = D_{KL}(q(\vartheta; \varphi) \,\|\, P(\vartheta)) - \mathbb{E}_q[\log P(o \mid \vartheta)]$$

$$= \text{complessità} - \text{accuratezza}$$

rivela la struttura del rasoio di Occam nella cognizione: il cervello cerca la spiegazione delle osservazioni più accurata con il minor aggiornamento del prior. La connessione con l'ELBO variazionale è esatta: F è l'ELBO negato, e le equazioni del predictive coding corrispondono al gradiente discendente su F rispetto ai parametri della distribuzione variazionale.

La teoria del predictive coding di Rao e Ballard (1999) specifica come le gerarchie corticali implementino questa minimizzazione: ogni strato mantiene rappresentazioni delle cause latenti e invia predizioni discendenti, ricevendo errori di predizione ascendenti. In un modello

Silvano Salvador

gerarchico a due livelli con errori $\varepsilon_0 = s - g_0(v^0)$ e $\varepsilon_1 = v^0 - g_1(v^1)$ e matrici di precisione Π_0, Π_1:

$$\partial F/\partial v^0 = -\Pi_0 \varepsilon_0 \cdot (\partial g_0/\partial v^0)^T + \Pi_1 \varepsilon_1$$

$$\partial F/\partial v^1 = -\Pi_1 \varepsilon_1 \cdot (\partial g_1/\partial v^1)^T$$

Le dinamiche di minimizzazione producono il flusso di errori di predizione ascendenti e discendenti attraverso la gerarchia corticale, con pesi dipendenti dalla precisione stimata a ogni livello — una modulazione che corrisponde formalmente all'attenzione selettiva.

Una valutazione intellettualmente onesta del principio di Friston deve però riconoscere la critica metodologica articolata da Colombo e Seriès (2012) e sviluppata sistematicamente da Colombo, Seriès e Sun (2021). Il loro argomento è che il principio nella sua formulazione più generale — ogni sistema che persiste nel tempo minimizza la propria energia libera — è non falsificabile nel senso di Popper: per qualsiasi sistema persistente è possibile costruire post hoc un modello generativo $P(o, \vartheta)$ e una distribuzione variazionale $q(\vartheta; \varphi)$ tali che il comportamento osservato corrisponda alla minimizzazione di F.

Friston ha risposto che il principio di energia libera non intende essere una teoria empirica nel senso popperiano, ma un framework normativo che descrive le condizioni di possibilità della cognizione adattiva. Questa risposta è filosoficamente coerente, ma ridimensiona drasticamente la portata esplicativa: un framework normativo compatibile con qualsiasi comportamento non spiega perché i sistemi cognitivi reali si comportino in un modo piuttosto che in un altro.

La connessione con i modelli variazionali profondi merita una precisazione analoga. La struttura matematica condivisa — $F = -\text{ELBO}$, decomposizione accuratezza-complessità, ottimizzazione variazionale —

Silvano Salvador

è reale e non banale. Ma la condivisione di una struttura matematica non implica la condivisione di meccanismi implementativi. Il cervello potrebbe realizzare l'inferenza variazionale attraverso dinamiche stocastiche di popolazione neurale che campionano dalla distribuzione posteriore, anziché attraverso l'ottimizzazione parametrica di una distribuzione approssimante gaussiana come nei VAE. Il rischio epistemologico è trattare l'isomorfismo formale come evidenza di un'architettura computazionale condivisa — un salto logico non giustificato dalla matematica del framework né dall'evidenza neuroscientifica disponibile.

Il valore scientifico del principio di Friston è quello di un'euristica generativa di grande fecondità: esso ha organizzato coerentemente osservazioni sperimentali sulla percezione, sull'attenzione e sulle patologie cognitive — in particolare la lettura della schizofrenia come errore nella stima della precisione degli errori di predizione — generando ipotesi testabili che ricerche successive hanno parzialmente confermato. Non è, per ora, una teoria quantitativa falsificabile nel senso in cui lo è la meccanica statistica o la termodinamica. Distinguere con rigore questi due livelli di validità epistemologica — euristica generativa e teoria quantitativa — è essenziale per valutare i progressi e i limiti della neuroscienza computazionale contemporanea.

L'informazione come fondamento — Parte Quarta

Compressione come intelligenza: MDL, leggi di scala e il paradosso della sovreparametrizzazione

Silvano Salvador

I. Il principio di lunghezza minima di descrizione

Il principio di lunghezza minima di descrizione (MDL) di Jorma Rissanen non è un semplice raffinamento tecnico del criterio di informazione di Akaike o del criterio bayesiano di Schwarz. È una riformulazione radicale della questione epistemologica al cuore dell'apprendimento statistico: che cosa significa scegliere razionalmente tra modelli concorrenti? La risposta è: il modello migliore è quello che permette la descrizione più corta dei dati. Non è una metafora computazionale del rasoio di Occam, ma la sua precisa formalizzazione operativa che si presta a derivazioni quantitative e a garanzie teoriche di approssimazione.

Il punto di partenza è la biunivocità tra distribuzioni di probabilità e schemi di codifica, conseguenza del teorema di codifica della sorgente di Shannon: ogni distribuzione p definisce uno schema in cui si assegna lunghezza $-\log_2 p(x)$ al simbolo x. Un modello che assegna alta probabilità ai dati è un modello che li comprime efficientemente; un modello che assegna bassa probabilità è verboso.

Il criterio MDL seleziona il modello M e i parametri θ che minimizzano la lunghezza totale:

$$L(D, \theta, M) = L(D \mid \theta, M) + L(\theta \mid M)$$

$$= -\log P(D \mid \theta, M) + L(\theta \mid M)$$

dove il primo termine è la log-verosimiglianza negativa e il secondo è la lunghezza necessaria per trasmettere i parametri con precisione sufficiente. Quando $L(\theta \mid M) = -\log P(\theta \mid M)$, questa struttura è identica all'inferenza bayesiana MAP: MDL e MAP sono lo stesso principio in vocabolari distinti.

II. Il codice NML e la complessità parametrica

Silvano Salvador

La normalized maximum likelihood (NML) è la forma più sofisticata del MDL ed evita la circolarità insita nella scelta della precisione di codifica dei parametri. Il codice NML assegna a ogni sequenza di dati x la distribuzione:

$$p_NML(x) = P(x \mid \hat{\theta}(x), M) \,/\, C(M)$$

dove $\hat{\theta}(x)$ è la stima di massima verosimiglianza per x e C(M) è la costante di normalizzazione:

$$C(M) = \sum_y P(y \mid \hat{\theta}(y), M)$$

La complessità parametrica C(M) misura il numero effettivo di distribuzioni distinte che M può generare. La lunghezza del codice NML per x è:

$$L_NML(x) = -\log P(x \mid \hat{\theta}(x), M) + \log C(M)$$

Il primo termine è la log-verosimiglianza massima negativa; il secondo è la complessità parametrica — una penalità indipendente da x che misura la flessibilità del modello in modo invariante rispetto alla parametrizzazione.

Per modelli regolari della famiglia esponenziale con d parametri e n dati, Rissanen (1996) ha dimostrato l'espansione asintotica:

$$\log C(M) = (d/2) \log(n/2\pi) + \log \int_\Theta \sqrt{\det G(\theta)}\, d\theta + O(1)$$

Il termine dominante $(d/2) \log n$ è identico alla penalità BIC di Schwarz, spiegando l'equivalenza asintotica tra BIC e MDL per modelli regolari. Il secondo termine — il logaritmo del volume riemanniano dello spazio dei parametri rispetto alla metrica di Fisher — è la correzione geometrica: misura quanto le diverse regioni di Θ contribuiscano ugualmente alla complessità.

Silvano Salvador

Il volume riemanniano $\int_\Theta \sqrt{\det G(\theta)}\, d\theta$ è la misura di Jeffreys sullo spazio dei parametri — il prior invariante per reparametrizzazioni che Jeffreys aveva proposto nel 1946 su basi filosofiche. Il MDL nella forma NML fornisce dunque una derivazione della misura di Jeffreys da principi puramente informazionali: è il prior che emerge naturalmente dalla struttura del problema di codifica, non da intuizioni soggettive sull'invarianza.

III. Il prior di Solomonoff e la non calcolabilità del limite teorico

Il prior di Solomonoff M(x) è la probabilità che un programma casuale estratto uniformemente bit a bit produca su una macchina di Turing universale U una sequenza che inizia con x:

$$M(x) = \sum_{p : U(p) = x^*} 2^{-|p|}$$

La connessione con la complessità di Kolmogorov K(x) — la lunghezza del programma più corto che produce x su U — è:

$$M(x) \geq 2^{-K(x)} \cdot c$$

per una costante $c > 0$, il che implica che le sequenze a bassa complessità di Kolmogorov ricevono alta probabilità. Il teorema di dominazione universale afferma che per ogni distribuzione semicalcolabile μ esiste c_μ tale che $M(x) \geq c_\mu \cdot \mu(x)$ per ogni x: un predittore basato su M non può essere sorpreso da nessuna distribuzione computabile.

La non calcolabilità di M non è soltanto una difficoltà pratica ma un ostacolo teorico fondamentale. La funzione K(x) non è computabile: non esiste algoritmo che, dato x, ne calcoli la complessità di Kolmogorov. Essa è soltanto semicalcolabile dall'alto — si possono costruire approssimazioni superiori — ma la convergenza non è garantita in

Silvano Salvador

tempo finito. Questa non calcolabilità ha una conseguenza metodologica di rilievo: l'affermazione che i modelli linguistici si avvicinino asintoticamente al predittore di Solomonoff è vera in senso operativo, ma la distanza da quel limite non può essere misurata. $K(\mu)$ è incalcolabile, e non è possibile sapere quanto si è lontani dall'ottimo teorico per nessuna distribuzione naturale specifica. Le leggi di scala documentano la traiettoria di avvicinamento, ma non possono quantificare la distanza residua dal limite.

VI-bis. Causal representation learning: oltre la correlazione statistica

Le leggi di scala e il principio MDL ottimizzano la compressione della distribuzione osservata: apprendono il modello più compatto compatibile con i dati di training. Questa ottimizzazione, per quanto potente, è fondamentalmente associativa: cattura correlazioni statistiche senza distinguere le relazioni causali che le sottendono. Schölkopf, Janzing, Peters e colleghi hanno sviluppato dal 2012 un programma di ricerca sul causal representation learning che si propone di apprendere rappresentazioni con struttura causale esplicita — rappresentazioni che identificano le variabili causali del dominio e le relazioni di causa-effetto tra di esse.

La distinzione formale tra correlazione e causa è catturata dal do-calculus di Pearl (2000). Mentre la distribuzione osservata $P(Y \mid X = x)$ risponde alla domanda "cosa osservo di Y quando osservo $X = x$?", la distribuzione interventionale $P(Y \mid do(X = x))$ risponde a "cosa osservo di Y quando intervengo a forzare $X = x$?". Le due distribuzioni coincidono solo quando X non ha cause comuni con Y non condizionate nell'intervento; in generale, $P(Y \mid X = x) \neq P(Y \mid do(X = x))$. Tutti i metodi di apprendimento automatico standard ottimizzano su

Silvano Salvador

P(Y | X), non su P(Y | do(X)), e quindi apprendono rappresentazioni che non generalizzano quando la distribuzione dei dati cambia per effetto di interventi.

Questo ha conseguenze dirette per il problema della distributional shift che attraversa l'intero saggio. Un modello addestrato su dati raccolti in condizioni A e deployato in condizioni B sovente fallisce non perché manchi di dati sufficienti, ma perché le condizioni B realizzano un intervento do(X = x) su variabili che nella distribuzione di training erano correlate con il target ma non causalmente connesse. Un classificatore medico addestrato su dati ospedalieri che usa la correlazione storica tra età e diagnosi come feature predittiva degrada quando cambia la popolazione ospedaliera — la correlazione è confounded dalla selezione demografica, non causale.

I metodi di Invariant Risk Minimization (IRM, Arjovsky et al., 2019) cercano rappresentazioni Φ tali che il predittore ottimale su Φ(X) sia lo stesso in tutti gli ambienti di training. Formalmente, IRM cerca Φ che minimizza:

$$\min_{\Phi, w} \sum_e L^e(w \cdot \Phi) + \lambda \cdot ||\nabla_w L^e(w \cdot \Phi)||^2$$

dove L^e è la perdita nell'ambiente e e il termine di penalizzazione forza il gradiente rispetto al predittore lineare w a essere piccolo — il che indica che w = 1 è già ottimale, e quindi Φ ha eliminato le features non invarianti attraverso gli ambienti. La rappresentazione risultante è invariante alle distribuzioni shift tra gli ambienti di training, il che è una condizione necessaria (ma non sufficiente) per la causalità.

La connessione con la compressione come intelligenza discussa nella Parte Quarta è profonda: il prior di Solomonoff favorisce le ipotesi più brevi, e le ipotesi causali — che specificano meccanismi generativi compatti — sono tipicamente più brevi delle ipotesi correlazionali —

Silvano Salvador

che richiedono di elencare tutte le correlazioni osservate. Un'intelligenza che ottimizza la compressione nel senso del prior di Solomonoff dovrebbe dunque preferire le strutture causali a quelle meramente correlazionali. Questa intuizione suggerisce che il causal representation learning non è un'estensione opzionale del paradigma informazionale ma una sua conseguenza necessaria: la compressione ottimale richiede la scoperta dei meccanismi causali sottostanti ai dati osservati.

IV. Le leggi di scala: struttura empirica e interpretazione teorica

Kaplan e colleghi (2020) hanno documentato che la perdita di un modello linguistico segue leggi di potenza nel numero di parametri N, nella quantità di dati D e nel budget computazionale C:

$$L(N) \approx (N_c / N)^{\alpha_N}, \quad \alpha_N \approx 0.076$$

$$L(D) \approx (D_c / D)^{\alpha_D}, \quad \alpha_D \approx 0.095$$

$$L(C) \approx (C_c / C)^{\alpha_C}, \quad \alpha_C \approx 0.050$$

Hoffmann e colleghi (2022) — il lavoro noto come Chinchilla — hanno derivato una legge congiunta più precisa:

$$L(N, D) \approx E + A / N^{\alpha} + B / D^{\beta}$$

dove $E \approx 1.69$ nat/token è la perdita irriducibile — l'entropia residua del linguaggio naturale non prevedibile da nessun modello — e A, B, α, β sono costanti empiriche. L'analisi di Chinchilla mostra che, dato un budget computazionale C misurato in FLOPs, la perdita è minimizzata dalla coppia ottimale:

$$N^* \propto C^{0.50}, \quad D^* \propto C^{0.50}$$

Silvano Salvador

ovvero dimensioni del modello e quantità di dati devono crescere con la radice quadrata del budget computazionale. Questa previsione — la legge di Chinchilla — implica che i modelli addestrati prima del 2022, incluso GPT-3 con 175 miliardi di parametri su 300 miliardi di token, fossero sistematicamente sottoaddetrati: il punto ottimale avrebbe richiesto circa 3.5 trilioni di token per quella dimensione.

L'interpretazione teorica delle leggi di scala in termini di criticalità fisica — l'analogia con i sistemi fisici alla criticalità dove le funzioni di correlazione decadono come leggi di potenza — merita cautela. Le leggi di scala osservate potrebbero riflettere la struttura degli ottimizzatori e degli obiettivi di addestramento, o la distribuzione statistica dei compiti nel corpus linguistico, senza implicare una transizione di fase nel senso del gruppo di rinormalizzazione. Ricerche successive hanno identificato deviazioni sistematiche: le emergent abilities documentate da Wei e colleghi (2022) — capacità qualitative che compaiono bruscamente oltre una certa soglia — non si lasciano descrivere da leggi di potenza continue. Tuttavia Schaeffer e colleghi (2023) hanno mostrato che molte di esse sono artefatti della metrica di valutazione: con metriche continue al posto di metriche discontinue, le transizioni apparentemente brusche diventano graduali, coerenti con le leggi di scala ordinarie. L'analogia con la criticalità fisica rimane dunque suggestiva ma non dimostrata.

V. La doppia discesa: struttura matematica e implicit bias

Il fenomeno della doppia discesa, documentato da Belkin e colleghi (2019), descrive la curva di errore di test in funzione della complessità del modello. Si osserva dapprima la curva a U classica nel regime sotto-parametrizzato, seguita da una seconda discesa nel regime sovra-parametrizzato, oltre la soglia di interpolazione dove il modello ha

Silvano Salvador

abbastanza parametri da interpolare esattamente ogni punto di addestramento.

Per la regressione lineare con n punti e p parametri con $p > n$, la discesa del gradiente converge alla soluzione di norma minima:

$$\hat{\beta}_{min} = X^T(XX^T)^{-1} y$$

Bartlett e colleghi (2020) hanno dimostrato che, sotto condizioni di coerenza bassa sulla struttura di X e distribuzione isotropa:

$$\mathbb{E}[L(\hat{\beta}_{min})] \leq \sigma^2 (n / (p-n)) + \|\beta^*\|^2 \cdot O(n/p)$$

Il primo termine diverge alla soglia di interpolazione $p = n$ — corrispondente al picco della doppia discesa — e decade come $n/(p-n)$ per $p \to \infty$. Il secondo termine decade come n/p: nel regime fortemente sovra-parametrizzato, la soluzione di norma minima generalizza perché la maggior parte della varianza del modello è concentrata nelle direzioni del segnale, con le direzioni ortogonali che contribuiscono con un termine piccolo.

L'implicit bias della discesa del gradiente — la tendenza a convergere verso la soluzione di norma minima tra quelle con errore di addestramento zero — è stato dimostrato rigorosamente per i modelli lineari (Soudry et al., 2018). Per le reti neurali profonde la situazione è più complessa: il bias implicito dipende dall'architettura, dall'ottimizzatore, dal tasso di apprendimento e dalla schedulazione, e non converge in generale alla soluzione di norma minima globale.

La connessione con il principio MDL è valida nella sua forma precisa soltanto per modelli lineari: la soluzione di norma minima è la più semplice compatibile con i dati nel senso della norma euclidea. Per le reti neurali profonde, la nozione di semplicità implicata dall'implicit bias della SGD non corrisponde alla complessità di Kolmogorov né alla lunghezza

Silvano Salvador

NML: dipende dalla norma del prodotto delle matrici di peso, dalla struttura gerarchica dell'architettura, dalla geometria del paesaggio della funzione di perdita. L'identità tra compressione e generalizzazione, valida per modelli lineari, diventa per le reti profonde un'analogia strutturale da trattare con la dovuta precisione.

VI. Il collo di bottiglia dell'informazione: portata e limiti

Il principio del collo di bottiglia dell'informazione di Tishby e Zaslavsky (2015) analizza ogni strato di una rete neurale come una soluzione di compromesso nel piano (I(X;T), I(T;Y)), dove T è la rappresentazione interna. L'obiettivo di ottimizzazione è:

$$\min_{p(t|x)} I(X; T) - \beta\, I(T; Y)$$

con $\beta > 0$ che controlla il compromesso tra compressione dell'input e conservazione dell'informazione rilevante. La soluzione ottimale appartiene alla curva del collo di bottiglia:

$$p^*(t|x) \propto p(t) \exp(\beta \sum_y p(y|x) \log[p(y|t)/p(y)])$$

un'equazione di punto fisso che generalizza la distribuzione di Gibbs al caso informazionale.

L'ipotesi sull'andamento dell'addestramento — una fase di fitting seguita da una fase di compressione in cui I(X;T) decresce — è stata contestata da Saxe e colleghi (2018), che hanno mostrato come la fase di compressione non si osservi nelle reti con attivazioni deterministiche (ReLU, tanh) ma soltanto con attivazioni che introducono stocasticità esplicita. La ragione formale: I(X;T) è invariante per trasformazioni invertibili di T, e se l'attivazione è iniettiva — come avviene per ReLU in molti regimi — T può essere ricostruito da X perfettamente, per cui I(X;T) = H(T) non varia. La compressione documentata da Tishby e

Silvano Salvador

Zaslavsky era un artefatto della stocasticità implicita in attivazioni sigmoidali saturanti.

Questo non invalida il framework concettuale del collo di bottiglia come criterio normativo per la qualità delle rappresentazioni. La domanda di quanto uno strato sia vicino alla frontiera Pareto ottimale nel piano $(I(X;T), I(T;Y))$ è ben posta e misurabile. Ma compromette la tesi più forte — che le reti neurali convergano durante l'addestramento verso rappresentazioni ottimali nel senso dell'information bottleneck — che rimane non dimostrata in generale.

VII. Compressione, intelligenza, certificabilità: il triangolo impossibile

Una caratteristica strutturale del paradigma informazionale che emerge dall'analisi congiunta del MDL, delle leggi di scala e del collo di bottiglia merita di essere esplicitata: la tensione tra compressione, prestazione e certificabilità è, per certi versi, irriducibile.

Un modello che comprime ottimalmente i dati nel senso del MDL è per definizione il più semplice compatibile con le osservazioni. Ma la semplicità nel senso di MDL — brevità della descrizione — non coincide con la trasparenza interpretativa, né con la certificabilità formale. Una rete neurale con miliardi di parametri che realizza una compressione ottimale di un corpus testuale è, in senso MDL, più semplice di un insieme di regole simboliche che si applicano a una sottoclasse di esempi, perché la rete comprime meglio. Ma non è né interpretabile né certificabile nel senso che i sistemi simbolici formalmente verificati lo sono.

Il sistema simbolico formalmente verificato offre garanzie di correttezza indipendenti dalla distribuzione dei dati di test: se la specifica formale del

Silvano Salvador

comportamento desiderato è corretta e la verifica è completa, il sistema si comporterà correttamente su ogni input nel dominio di specifica. Il sistema neurale addestrato offre garanzie probabilistiche condizionate alla distribuzione di test: se la distribuzione di test è simile a quella di addestramento e la funzione di perdita è allineata con l'obiettivo reale, allora il sistema si comporterà probabilmente bene su input tipici. Il condizionale debole che qualifica le garanzie del paradigma informazionale non è un difetto tecnico superabile: è la conseguenza strutturale dell'apprendimento da distribuzione finita senza specifica formale del comportamento desiderato.

Le applicazioni dove la certificabilità è non negoziabile — avionica, dispositivi medici impiantabili, sicurezza crittografica, infrastrutture critiche — richiedono dunque architetture ibride che combinano il paradigma informazionale per la percezione e il pattern matching con il paradigma simbolico formale per la specifica e la verifica del comportamento. Non si tratta di un compromesso ingegneristico tra tradizioni concorrenti: è la risposta tecnicamente corretta alla struttura del problema.

VIII. Il problema dell'allineamento come limite interno al paradigma

La monografia rinvia al terzo saggio per la trattazione sistematica del problema dell'allineamento, ma una sua menzione è necessaria già qui perché esso emerge come limite interno alla struttura stessa dell'apprendimento statistico, non come questione applicativa separabile.

Un sistema addestrato a minimizzare una funzione di perdita su un corpus finito apprende a ottimizzare quella funzione di perdita su quel corpus — non necessariamente a perseguire l'obiettivo che la funzione di perdita intendeva approssimare. La specificazione della funzione di

Silvano Salvador

perdita è un atto interpretativo che introduce un gap tra la rappresentazione formale dell'obiettivo e l'obiettivo reale. Goodhart aveva formulato questo problema come legge generale nel 1975: quando una misura diventa un obiettivo, cessa di essere una buona misura.

Nel Reinforcement Learning from Human Feedback (RLHF), la funzione di ricompensa appresa dai feedback umani è un modello della preferenza umana media sulla distribuzione di prompt su cui è stata addestrata — non una rappresentazione delle preferenze umane in generale. Il sistema addestrato a massimizzare questa funzione include la possibilità di trovare comportamenti che ottengono alta ricompensa dal modello di preferenza senza corrispondere alle preferenze reali: il reward hacking. Questa non è una difficoltà tecnica correggibile con più dati di feedback: è la manifestazione del teorema di Goodhart nell'apprendimento statistico.

La connessione con la struttura informazionale del paradigma è diretta: il mismatch tra la distribuzione di ricompensa appresa e quella reale è una forma di misspecification del modello di preferenza, analoga alla misspecification discussa nella Parte Seconda. La soluzione richiede non soltanto migliori dati di preferenza, ma probabilmente una revisione strutturale dell'approccio che combini specifiche formali del comportamento desiderato con apprendimento statistico delle preferenze — un problema aperto che il prossimo saggio affronterà nelle sue articolazioni tecniche.

L'informazione come fondamento — Parte Quinta

Silvano Salvador

Modelli generativi come ontologia probabilistica: GAN, trasporto ottimale e la questione del generare come comprendere

I. Il problema della generazione come problema di stima di distribuzione

Ogni modello generativo affronta, nella sua essenza, un unico problema: data una distribuzione empirica P_data nota soltanto attraverso campioni finiti $\{x_1, \ldots, x\}$, costruire una distribuzione parametrica P_model che ne approssimi la struttura in modo da poter generare nuovi campioni di alta qualità. Le famiglie di modelli generativi profondi si differenziano per la scelta di quale funzionale della coppia (P_data, P_model) ottimizzare e per le soluzioni architetturali adottate per rendere questo obiettivo computazionalmente trattabile.

I modelli variazionali della Parte Seconda ottimizzano l'ELBO — un lower bound alla log-verosimiglianza — attraverso un'architettura encoder-decoder che rende l'inferenza sui latenti computazionalmente accessibile. I modelli di diffusione della Parte Terza ottimizzano la stima della score function attraverso un processo di aggiunta e rimozione progressiva di rumore. I normalizing flows, discussi in questa parte e nell'Appendice III, ottimizzano la log-verosimiglianza esatta attraverso trasformazioni invertibili con jacobiano trattabile. Le reti antagoniste generative, il cui framework è descritto nelle sezioni seguenti, ottimizzano invece una divergenza tra distribuzioni attraverso una competizione tra due reti. Ciascun approccio comporta compromessi distinti tra fedeltà della densità appresa, efficienza della generazione, espressività del modello e garanzie teoriche sulla qualità dell'approssimazione.

Silvano Salvador

II. Le reti antagoniste generative: il gioco minimax e la divergenza JS

Le generative adversarial networks di Goodfellow e colleghi (2014) costituiscono l'approccio alla generazione profonda più distante dal paradigma bayesiano variazionale. Il gioco minimax che le definisce è:

$$\min_G \max_D V(D, G) = \mathbb{E}_{x \sim P_{data}}[\log D(x)] + \mathbb{E}_{z \sim P_z}[\log(1 - D(G(z)))]$$

dove $D : \mathbb{R}^n \to [0,1]$ è il discriminatore — che apprende a distinguere campioni reali da sintetici — e $G : \mathbb{R}^k \to \mathbb{R}^n$ è il generatore — che mappa rumore latente $z \sim P_z(z)$ in campioni realistici.

Per D ottimale con G fissato, la derivata dell'obiettivo rispetto a D(x) produce, per ogni x:

$$D^*_G(x) = P_{data}(x) / (P_{data}(x) + P_G(x))$$

Sostituendo D^*_G nell'obiettivo del generatore:

$$V(D^*_G, G) = \mathbb{E}_{x \sim P_{data}}[\log(P_{data}(x)/(P_{data}(x)+P_G(x)))]$$

$$+ \mathbb{E}_{x \sim P_G}[\log(P_G(x)/(P_{data}(x)+P_G(x)))]$$

$$= 2\, D_{JS}(P_{data} \| P_G) - \log 4$$

dove la divergenza di Jensen-Shannon è:

$$D_{JS}(P \| Q) = \tfrac{1}{2} D_{KL}(P \| M) + \tfrac{1}{2} D_{KL}(Q \| M), \quad M = \tfrac{1}{2}(P + Q)$$

La divergenza JS è simmetrica e limitata superiormente da log 2. Il minimo dell'obiettivo del generatore è raggiunto quando $P_G = P_{data}$.

La derivazione è formalmente impeccabile nel contesto del discriminatore ottimale. In pratica il discriminatore non è mai ottimale: è una rete addestrata con gradiente stocastico che alterna aggiornamenti

Silvano Salvador

con il generatore. La non-stazionarietà del paesaggio di perdita per il generatore produce dinamiche che non corrispondono a nessuna traiettoria di discesa del gradiente su una funzione scalare fissa.

La radice analitica dell'instabilità è identificabile con precisione. Quando i supporti di P_data e P_G sono disgiunti — circostanza quasi inevitabile nelle fasi iniziali, quando il generatore produce campioni lontani dalla distribuzione reale — la divergenza JS vale identicamente log 2 indipendentemente dalla distanza tra le distribuzioni. Il suo gradiente rispetto ai parametri di G è dunque identicamente zero: il generatore non riceve alcun segnale informativo su come muoversi verso P_data. Questo non è un difetto tecnico correggibile con architetture più sofisticate: è una proprietà analitica della divergenza JS, connessa al fatto che essa misura la distinguibilità locale tra distribuzioni e non la loro distanza globale. La soluzione deve dunque provenire da una sostituzione della divergenza stessa.

III. La distanza di Wasserstein: geometria del trasporto ottimale

La soluzione proposta da Arjovsky, Chintala e Bottou (2017) con le Wasserstein GAN consiste nel sostituire la divergenza JS con la distanza di Wasserstein-1, o Earth Mover's Distance. Per due distribuzioni P e Q su uno spazio metrico ($\mathscr{X}$, d), questa distanza è:

$$W_1(P, Q) = \inf_{\gamma \in \Pi(P,Q)} \mathbb{E}_{(x,y) \sim \gamma}[d(x, y)]$$

dove $\Pi(P,Q)$ è l'insieme di tutti i piani di trasporto — distribuzioni di probabilità su $\mathscr{X} \times \mathscr{X}$ con marginali P e Q rispettivamente. La formulazione duale di Kantorovich-Rubinstein afferma che, per d metrica euclidea:

$$W_1(P, Q) = \sup_{\|f\|_L \leq 1} [\mathbb{E}_{x \sim P}[f(x)] - \mathbb{E}_{y \sim Q}[f(y)]]$$

Silvano Salvador

dove il supremo è sull'insieme delle funzioni 1-Lipschitz $f : \mathbb{R}^n \to \mathbb{R}$. La derivazione si basa sulla dualità di programmazione lineare per il problema di Kantorovich. Il problema primale è:

$\min_{\gamma \geq 0} \iint d(x,y)\, d\gamma(x,y)$ soggetto a marginali P e Q

Il suo duale è:

$\max_{f,g} \int f\, dP + \int g\, dQ$ soggetto a $f(x) + g(y) \leq d(x,y)$

Con la scelta ottimale $g(y) = -f(y)$, il vincolo diventa $f(x) - f(y) \leq d(x,y)$, ovvero $\|f\|_L \leq 1$. La dualità forte assicura che il valore ottimo primale e duale coincidano, giustificando la formulazione di Kantorovich-Rubinstein.

La superiorità di W_1 sulla divergenza JS per l'addestramento generativo è connessa alla continuità rispetto alla convergenza debole delle distribuzioni. Per qualsiasi sequenza $P_n \to P$ debolmente, si ha $W_1(P_n, Q) \to W_1(P, Q)$. La divergenza JS non gode di questa proprietà quando le distribuzioni hanno supporti diversi: è discontinua e non fornisce gradiente informativo nel caso tipico dei supporti disgiunti.

Nell'implementazione delle WGAN il vincolo di Lipschitz sul critico f viene imposto pragmaticamente. Il clipping dei pesi della versione originale di Arjovsky et al. è un'approssimazione grossolana: non garantisce la condizione 1-Lipschitz e richiede la regolazione manuale della soglia c. La gradient penalty di Gulrajani e colleghi (2017) è più principiata: si aggiunge all'obiettivo il termine

$\lambda\, \mathbb{E}_{\hat{x} \sim p_{\hat{x}}}[(\|\nabla_{\hat{x}} f(\hat{x})\|_2 - 1)^2]$

dove $\hat{x} = \varepsilon\, x + (1-\varepsilon)\, G(z)$ con $\varepsilon \sim U(0,1)$. Il gradiente di una funzione 1-Lipschitz ha norma al più 1 ovunque e, all'ottimo, norma esattamente 1 sui percorsi geodetici tra P e Q. La penalità approssima il vincolo globale

Silvano Salvador

concentrandosi sui punti interpolati, che sono quelli più informativi per il critico durante l'addestramento.

IV. Normalizing flows: la fedeltà della densità e i suoi costi architetturali

I normalizing flows massimizzano la precisione analitica a scapito della flessibilità architettonica. La log-densità di x = f_K ◦ … ◦ f_1(z) con z ~ P_z si calcola esattamente:

$$\log p(x) = \log p_z(z) - \sum_{k=1}^{K} \log |\det J_{f_k}(z_{k-1})|$$

Il calcolo del determinante jacobiano è in generale $O(n^3)$, rendendolo proibitivo per dati ad alta dimensionalità. Le architetture pratiche sono progettate per rendere questo calcolo O(n).

Le coupling layers di RealNVP suddividono x in (x_A, x_B) e applicano:

$$y_A = x_A, \quad y_B = x_B \odot \exp(s(x_A)) + t(x_A)$$

La jacobiana è triangolare a blocchi con determinante:

$$\log |\det J_f| = \sum_i s_i(x_A) = 1^T s(x_A)$$

calcolabile in O(d) indipendentemente dalla complessità delle reti s e t. L'inversione è esplicita: $x_A = y_A$, $x_B = (y_B - t(y_A)) \odot \exp(-s(y_A))$.

Le autoregressive flows sfruttano la fattorizzazione $p(x) = \prod_i p(x_i \mid x_{<i})$ con trasformazioni $x_i = \mu_i(x_{<i}) + \sigma_i(x_{<i}) \cdot \varepsilon_i$. La jacobiana è triangolare inferiore con log-determinante:

$$\log |\det J| = \sum_i \log \sigma_i(x_{<i})$$

calcolabile in O(d), ma la generazione richiede n passi sequenziali. MAF (Masked Autoregressive Flow) è efficiente per la stima della densità ma lento nella generazione; IAF (Inverse Autoregressive Flow) è efficiente

Silvano Salvador

nella generazione ma lento nella stima della densità: le due architetture sono duali rispetto a questa coppia di requisiti.

Glow introduce le convoluzioni 1×1 invertibili con decomposizione LU: $W = PLU$ con P permutazione fissa, L triangolare inferiore con diagonale unitaria, U triangolare superiore. Il log-determinante è $\sum_i \log |U_{ii}|$, calcolabile in $O(C)$ invece che $O(C^3)$, riducendo il costo dominante nella stima della densità su tensori con molti canali. La trattazione tecnica completa delle architetture di normalizing flows è sviluppata nell'Appendice III.

La tensione fondamentale dei flows — tra fedeltà analitica della densità e vincoli imposti dall'invertibilità — non è risolvibile all'interno del paradigma: l'invertibilità è condizione necessaria per il teorema del cambiamento di variabile, e non può essere rilassata senza perdere la log-verosimiglianza esatta. La classe di trasformazioni invertibili con jacobiano trattabile è sufficientemente restrittiva da limitare l'espressività del modello rispetto ad architetture senza vincoli di invertibilità — spiegando perché i flows non abbiano raggiunto la qualità generativa dei modelli di diffusione nelle applicazioni pratiche ad alta risoluzione, nonostante le superiori garanzie di densità.

V. La semantica distribuzionale e la geometria dello spazio latente

La semantica distribuzionale contemporanea — nella sua versione dei word embeddings e dei transformer — può essere inquadrata formalmente attraverso la nozione di informazione mutua puntuale (PMI). Per un corpus, la PMI tra una parola w e un contesto c è:

$$\mathrm{PMI}(w, c) = \log \left[P(w, c) \,/\, (P(w)\, P(c)) \right]$$

Silvano Salvador

Levy e Goldberg (2014) hanno dimostrato che il modello word2vec nella variante skip-gram fattorizza implicitamente la matrice di PMI spostata:

$$W \cdot C \approx PMI(w, c) - \log k$$

dove W e C sono le matrici di embedding delle parole e dei contesti e k è il numero di campioni negativi. Il vettore di embedding è la rappresentazione a bassa dimensionalità della riga corrispondente della matrice di PMI — una compressione dell'informazione mutua tra la parola e i suoi possibili contesti. In questo senso il significato distribuzionale di una parola è la firma delle sue co-occorrenze statistiche: la massima di Firth — «you shall know a word by the company it keeps» — riceve una formalizzazione informazionale precisa.

I transformer estendono questo principio con rappresentazioni contestuali: il vettore di una parola dipende dall'intero contesto in cui occorre. Il meccanismo di attenzione:

$$\text{Attention}(Q, K, V) = \text{softmax}(QK^T/\sqrt{d_k})\, V$$

realizza — come notato nella Parte Terza — un recupero dalla memoria contestuale pesato per similarità. Il vettore contestuale di una parola è una media pesata dei valori nel contesto corrente, con pesi proporzionali alla similarità tra la query della parola e le chiavi dei contesti. Il significato contestuale emerge come distribuzione continua su possibili significati, pesata dall'informazione mutua con il contesto: un'implementazione del principio di Firth che è al tempo stesso un'operazione di recupero bayesiano dalla memoria contestuale.

La geometria dello spazio latente appresa dai modelli generativi condizionati ha una proprietà notevole ma sottovalutata: essa è continua rispetto alle categorie cognitive umane, che sono invece discrete. Un modello generativo di volti apprende direzioni nello spazio latente

Silvano Salvador

corrispondenti a età, orientazione, espressione — assi continui di variazione — non etichette discrete. Questa struttura continua è epistemicamente onesta: le proprietà dei domini naturali variano in modo continuo, e la discretizzazione categoriale è un'operazione cognitiva sovrapposta alla struttura del mondo, non una proprietà del mondo stesso.

Il rovescio di questa continuità è il disallineamento tra lo spazio latente del modello e lo spazio linguistico dell'utente. Quando si vuole generare «una persona di mezza età con occhiali rotondi», l'istruzione deve essere tradotta in una posizione nello spazio latente che non corrisponde a coordinate canoniche: lo spazio latente non ha assi etichettati, e i fattori di variazione appresi non corrispondono necessariamente alle categorie linguistiche dell'utente. Il modello CLIP di Radford e colleghi (2021) tenta di ridurre questo disallineamento apprendendo uno spazio latente comune testo-immagine attraverso un obiettivo contrastivo. Ma la struttura di tale spazio dipende dalla distribuzione delle coppie testo-immagine nel corpus, che riflette le categorie e le pratiche descrittive della comunità produttrice di quel corpus — non uno spazio semantico neutrale o universale. Un modello addestrato prevalentemente su testi e immagini di comunità anglofone occidentali apprende un spazio latente che generalizza in modo imprevedibile a domini culturali diversi.

VI. Il problema dell'allucinazione come mismatch distribuzionale multiplo

Le allucinazioni — generazione di affermazioni false con apparente confidenza — sono presentate spesso come un difetto tecnico correggibile. Un'analisi strutturale rivela invece che esse emergono da più

Silvano Salvador

cause distinte con strutture diverse, ciascuna richiedente soluzioni di tipo differente.

La prima causa è il mismatch di corpus. I testi disponibili su scala web non sono un campione imparziale delle affermazioni vere sul mondo: contengono errori, credenze false, narrazioni fittizie, ironie, satire. Un modello addestrato a prevedere il prossimo token su questo corpus apprende la distribuzione dei testi — non la distribuzione delle verità. Quando genera un'affermazione storicamente falsa ma stilisticamente comune nel web, non sta sbagliando rispetto alla propria distribuzione appresa: sta producendo un campione ad alta probabilità da quella distribuzione. Il mismatch è tra la distribuzione del corpus e la distribuzione delle verità.

La seconda causa è strutturale alla predizione autoregressiva. Il modello massimizza la coerenza locale — ogni token è probabile dato il contesto immediato — senza garantire la coerenza globale su scale lunghe. Una sequenza può essere localmente coerente in ogni punto eppure contenere contraddizioni a distanza, perché la distribuzione condizionale locale non incorpora vincoli di consistenza globale.

La terza causa riguarda la distribuzione dei fatti rari. I fatti che compaiono raramente nel corpus sono rappresentati con bassa frequenza nei dati di addestramento. La stima della loro probabilità di verità è intrinsecamente più incerta, ma il modello non ha accesso a questa gerarchia di certezze: il segnale di addestramento tratta ogni token con lo stesso obiettivo di predizione. Il risultato è che il modello genera fatti rari con la stessa apparente confidenza dei fatti comuni.

Queste tre cause distinte richiedono soluzioni parzialmente diverse. Al mismatch di corpus risponde la cura dei dati e l'RLHF. All'incoerenza globale risponde il retrieval-augmented generation. All'incertezza sui fatti

Silvano Salvador

rari risponde la calibrazione dell'incertezza e la generazione condizionata a segnali di confidenza espliciti. Nessuna soluzione è completa in isolamento, e la loro integrazione rimane un problema aperto della ricerca sull'allineamento.

VII. La questione filosofica: generare come comprendere

La questione più profonda che i modelli generativi pongono non è tecnica ma filosofica: la capacità di generare artefatti di alta qualità in un dominio implica una forma di comprensione di quel dominio? La risposta a questa domanda ha conseguenze pratiche immediate — su come calibrare la fiducia nelle previsioni dei modelli, su come interpretarne i fallimenti, su come strutturare la supervisione umana — ed è dunque una questione a cui non ci si può sottrarre con un rinvio alla filosofia speculativa.

L'argomento della stanza cinese di Searle (1980) afferma che la manipolazione sintattica di simboli non produce comprensione semantica, indipendentemente dalla complessità della manipolazione. Il ragionamento è il seguente: un agente che non comprende il cinese esegue un programma di manipolazione di simboli cinesi secondo regole sintattiche esplicite; l'agente produce risposte indistinguibili da quelle di un parlante nativo; ciononostante l'agente non comprende il cinese, perché non ha accesso al significato dei simboli che manipola. Se questo è concesso, nessun sistema computazionale può comprendere in virtù della sola computazione sintattica.

La risposta wittgensteiniana — il significato è uso, e un sistema che usa le parole nei modi tipici manifesta una padronanza semantica nel senso in cui la padronanza semantica è attribuibile — è potente ma non risolutiva per una ragione precisa che vale articolare. Wittgenstein sosteneva che il significato è determinato dall'uso nelle pratiche condivise di una

Silvano Salvador

comunità linguistica. Ma un modello linguistico non partecipa a pratiche condivise nel senso wittgensteiniano: non agisce nel mondo, non ha conseguenze delle proprie asserzioni, non può essere tenuto responsabile di ciò che afferma. La sua relazione con il linguaggio è essenzialmente passiva — un riflesso statistico delle pratiche altrui — non la partecipazione attiva a una forma di vita che Wittgenstein considerava costitutiva del significato.

L'obiezione non intende concludere che i modelli linguistici non abbiano alcuna forma di comprensione. Intende chiarire che la risposta wittgensteiniana sposta il problema senza risolverlo: la domanda diventa se la partecipazione attiva alle pratiche sia necessaria per il significato, o se un riflesso statistico sufficientemente accurato sia sufficiente. Su questo punto Wittgenstein non offre una risposta univoca, e la questione rimane aperta.

Il prossimo saggio riprende questa questione nella sezione conclusiva dedicata a Wittgenstein e al linguaggio dell'intelligenza, a partire dai risultati sul comportamento degli agenti ottimizzanti e sul problema della specificazione dei valori. Il lettore troverà lì un approccio complementare: dove questa sezione interroga il significato del generare in relazione all'uso linguistico, il Saggio III interroga il significato del comprendere in relazione all'agire nel mondo — la distinzione tra competenza linguistica e agentività intenzionale che Wittgenstein considerava inseparabili.

Tre posizioni distinte sono difendibili. La prima è il deflazionismo semantico: la comprensione è operazionalmente definita dalla capacità di usare il linguaggio in modo appropriato al contesto, e i modelli linguistici che soddisfano questo criterio comprendono nel senso operazionale del termine. La seconda è il realismo semantico: la comprensione richiede

Silvano Salvador

connessione causale con il mondo — intenzionalità nel senso di Brentano — che i modelli linguistici non hanno; la loro competenza linguistica è sintassi sofisticata senza semantica. La terza è il gradualismo funzionale: la comprensione è un continuum di capacità funzionali, e i modelli linguistici si trovano su questo continuum in posizione avanzata rispetto ai sistemi simbolici classici, senza che si possa tracciare un confine netto tra comprensione autentica e sua simulazione.

Il presente saggio non prende posizione definitiva tra queste tre opzioni — non perché la questione sia irrilevante, ma perché le evidenze disponibili non la risolvono. Ciò che si può affermare con maggiore sicurezza è questo: la capacità generativa dei modelli profondi è connessa a strutture rappresentazionali di una complessità e ricchezza che non si erano mai raggiunte con i paradigmi precedenti. Che questa complessità costituisca o meno comprensione nel senso filosoficamente robusto del termine è una questione empiricamente aperta che i progressi dell'interpretabilità meccanicistica degli anni prossimi potranno illuminare, non risolvere definitivamente.

VIII. L'AI neuro-simbolica e il paradigma ibrido

L'analisi dei modelli generativi come ontologia probabilistica sarebbe incompleta senza considerare la linea di sviluppo neuro-simbolica, che non è semplicemente un compromesso ingegneristico tra due tradizioni, ma un programma di ricerca con fondamenti propri.

Il benchmark SCAN di Lake e Baroni (2018) ha mostrato che le reti seq2seq standard falliscono su compiti di generalizzazione composizionale che qualsiasi bambino di cinque anni risolve: se il modello ha appreso il significato di «jump» e di «twice», dovrebbe generalizzare a «jump twice» senza training esplicito su quella combinazione. Le reti standard non generalizzano in questo modo. Il

Silvano Salvador

Compositional Generalization Problem è strutturale: le reti neurali apprendono associazioni statistiche tra pattern, non regole composizionali con applicazione sistematica.

I grandi modelli linguistici contemporanei mostrano capacità composizionali emergenti che parzialmente affrontano questo problema — ma la loro generalizzazione composizionale dipende dalla presenza del pattern composizionale nel corpus di addestramento, non da una struttura interna che garantisce la composizionalità come proprietà formale. La distinzione è cruciale: la generalizzazione basata su distribuzione statistica è fragile fuori distribuzione; quella basata su struttura formale è robusta per costruzione.

L'architettura neuro-simbolica tipica combina un modulo neurale — responsabile della percezione e del pattern matching su input non strutturati — con un modulo simbolico — responsabile del ragionamento strutturato, della pianificazione e della verifica di vincoli. Questo accoppiamento risponde alla distinzione cognitiva tra il ragionamento tipo-1 — rapido, parallelo, associativo — e il ragionamento tipo-2 — lento, sequenziale, governato da regole — che la psicologia cognitiva ha documentato estensivamente. Il paradigma informazionale è ottimale per il tipo-1; il paradigma simbolico per il tipo-2. Un'intelligenza artificiale robusta potrebbe richiedere entrambi.

L'architettura neuro-simbolica tipica combina un modulo neurale con un modulo simbolico attraverso quattro pattern di integrazione distinti. Il primo pattern è la percezione neurale con ragionamento simbolico: la rete neurale processa gli input grezzi e produce rappresentazioni intermedie passate a un reasoner simbolico (sistema di inferenza logica, planner classico, SAT solver). AlphaGo nella versione originale realizza questo pattern: la rete neurale stima probabilità di mossa e valore della

Silvano Salvador

posizione, mentre il Monte Carlo Tree Search le usa come euristiche per una ricerca esplicita nello spazio degli stati.

Il secondo pattern è l'apprendimento guidato da prior strutturati. DreamCoder (Ellis et al., 2021) impara librerie di programmi riusabili attraverso un ciclo wake-sleep ispirato all'inferenza bayesiana: nella fase wake usa la libreria corrente per risolvere task; nella fase sleep genera nuovi programmi che comprimono le soluzioni e aggiorna la libreria. Questo ciclo realizza il principio MDL della Parte Quarta: la libreria converge verso la rappresentazione che minimizza la lunghezza di descrizione dei task osservati.

Il terzo pattern è la differenziazione attraverso regole: Graph Neural Networks e Neural Logic Machines incorporano operazioni relazionali differenziabili che apprendono end-to-end rappresentazioni con struttura simbolica implicita. La struttura relazionale emerge come bias architetturale, non come modulo separato, ed è quindi non direttamente ispezionabile come regole enumerabili.

Il quarto pattern è il ragionamento neurale multi-passo attraverso chain-of-thought. I grandi modelli linguistici come GPT-4 e Claude mostrano capacità composizionali emergenti quando esplicitano i passi intermedi del ragionamento (Wei et al., 2022). Il testo del ragionamento intermedio funge da working memory simbolica che struttura la computazione in passi sequenziali, aggirando i limiti di composizionalità del forward pass singolo. Il programma synthesis, come FlashFill di Gulwani (2011), realizza il caso limite: combinando DSL simbolico e ricerca neuralmente guidata, produce programmi corretti da esempi di input-output in domini strutturati.

La valutazione complessiva è che il paradigma neuro-simbolico risponde correttamente alla diagnosi del problema — i sistemi puramente statistici

Silvano Salvador

mostrano difficoltà con composizionalità sistematica e ragionamento strutturato — ma non ha ancora prodotto una soluzione generale che scala come i modelli puramente neurali. I grandi modelli linguistici contemporanei manifestano capacità neuro-simboliche emergenti attraverso chain-of-thought e tool use senza architettura ibrida esplicita; resta aperta la questione di se questa emergenza sia robusta e generalizzabile.

IX. Il meccanismo di self-attention e l'architettura transformer: fondamenti informativi

I modelli di diffusione, i VAE e le GAN che le sezioni precedenti hanno analizzato come implementazioni del principio informazionale condividono una limitazione architetturale: la loro capacità di modellare dipendenze a lungo raggio nelle sequenze è limitata dalla topologia locale delle convoluzioni o dalla struttura delle connessioni ricorrenti. Il meccanismo di self-attention, introdotto da Vaswani e colleghi nel lavoro fondativo "Attention is All You Need" (2017), risolve questa limitazione con un'architettura che può essere interpretata direttamente come operazione informazionale: il calcolo di una distribuzione di probabilità contestuale che determina quali informazioni estrarre da quali posizioni.

Formalmente, dato un input $X \in \mathbb{R}^{n \times d}$ di n token con dimensione d, il meccanismo di self-attention proietta X attraverso tre matrici di pesi $W_Q, W_K, W_V \in \mathbb{R}^{d \times d_k}$ per produrre le matrici di query, chiave e valore:

$$Q = XW_Q, \quad K = XW_K, \quad V = XW_V$$

e calcola l'output come:

$$\text{Attention}(Q, K, V) = \text{softmax}(QK^T / \sqrt{d_k}) \cdot V$$

Silvano Salvador

Il termine QK^T è la matrice delle somiglianze tra ogni coppia di posizioni (query, chiave), scalata per √d_k per prevenire la saturazione della softmax in alta dimensione. La softmax trasforma queste somiglianze in una distribuzione di probabilità $A \in \mathbb{R}^{n\times n}$ dove A_{ij} è l'attenzione che la posizione i presta alla posizione j. L'output è la media pesata dei vettori valore V secondo questa distribuzione. Ogni riga i di A è quindi una distribuzione di probabilità sulle posizioni, e il token output alla posizione i è la media attesa dei valori sotto quella distribuzione.

L'interpretazione informazionale è diretta. La distribuzione $A_{i\cdot}$ è proporzionale a $\exp(q_i \cdot k_j / \sqrt{d_k})$: è una distribuzione di Gibbs con energia $E_{ij} = -q_i \cdot k_j / \sqrt{d_k}$ e temperatura $\beta = 1$. Come discusso nella connessione tra funzione di partizione e massima entropia nella Parte Prima, minimizzare l'energia corrisponde a selezionare le posizioni più "simili" alla query corrente, mentre la temperatura √d_k controlla la nitidezza della distribuzione: con valori di temperatura bassi, la distribuzione concentra tutta l'attenzione sulla posizione più simile; con valori alti, l'attenzione si distribuisce uniformemente. Il meccanismo di scaling √d_k implementa quindi un controllo entropico dell'attenzione, mantenendo l'entropia della distribuzione $A_{i\cdot}$ nell'ordine di grandezza di 1 nit indipendentemente dalla dimensionalità d_k.

La multi-head attention esegue h istanze parallele del meccanismo, proiettando Q, K, V con matrici distinte per ciascuna testa:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, ..., \text{head}_h)\, W_O, \quad \text{dove head}_i = \text{Attention}(XW_Q^i, XW_K^i, XW_V^i)$$

Le multiple teste apprendono distribuzione di attenzione con caratteristiche diverse: alcune si specializzano in relazioni sintattiche locali, altre in coreferenze a lungo raggio, altre in relazioni semantiche specifiche del dominio. Questa specializzazione è analoga alla

Silvano Salvador

decomposizione spettrale di una matrice di informazione mutua: ciascuna testa cattura una "dimensione di dipendenza" distinta tra le posizioni, e la concatenazione le combina in una rappresentazione multidimensionale.

Un blocco transformer completo combina la multi-head attention con una rete feed-forward a due strati (FFN), entrambi con connessioni residuali e normalizzazione dei layer:

$$x' = \text{LayerNorm}(x + \text{MultiHead}(x, x, x)), \quad \text{output} = \text{LayerNorm}(x' + \text{FFN}(x'))$$

Il ruolo informazionale del blocco FFN non è irrilevante: ricerche recenti sull'interpretabilità meccanicistica hanno mostrato che i blocchi FFN fungono da "memorie chiave-valore" (Geva et al., 2021) dove le attivazioni di primo strato corrispondono a pattern (chiavi) e le colonne di secondo strato corrispondono a distribuzioni di output (valori). Questa struttura è formalmente equivalente a un meccanismo di recupero da memoria associativa, connettendo l'architettura transformer alle Hopfield networks modernizzate discusse nella Parte Terza.

La connessione con le leggi di scala discusse nella Parte Quarta è strutturale. L'efficacia del transformer nell'apprendimento linguistico scala come legge di potenza con il numero di parametri N, la dimensione del dataset D e il compute C, con esponenti che suggeriscono una convergenza asintotica verso il prior di Solomonoff. GPT-4, Claude e i modelli contemporanei più capaci sono transformer autoregressivi addestrati con le leggi di scala come guida progettuale: la comprensione di questa architettura come implementazione computazionale del principio informazionale è quindi non soltanto storicamente illuminante ma praticamente indispensabile per interpretare lo stato dell'arte nell'intelligenza artificiale applicata.

Silvano Salvador

X. Il contrastive learning e l'apprendimento di rappresentazioni multimodali

I modelli generativi discussi nelle sezioni precedenti apprendono rappresentazioni latenti come effetto collaterale del task di generazione: il VAE impara uno spazio z che permette la ricostruzione fedele di x, il modello di diffusione impara uno score che guida il denoising. Una linea di ricerca parallela — il contrastive learning — apprende rappresentazioni imponendo direttamente una struttura informazionale sullo spazio latente: le rappresentazioni di esempi "simili" devono essere vicine, quelle di esempi "dissimili" devono essere lontane. Questa struttura può essere formalizzata come massimizzazione dell'informazione mutua tra le rappresentazioni di coppie di esempi correlati.

CLIP (Contrastive Language-Image Pre-training, Radford et al., 2021) è l'implementazione più influente di questo principio nel regime multimodale. Dati N coppie (immagine_i, testo_i) di esempi correlati (un'immagine e la sua didascalia), CLIP addestra due encoder — uno visivo f_v e uno testuale f_t — massimizzando la somiglianza coseno tra le coppie corrette (i, i) e minimizzandola tra le coppie scorrette (i, j), $j \neq i$. La perdita InfoNCE, introdotta da van den Oord et al. (2018), formalizza questo obiettivo:

$$L_{CLIP} = -(1/N) \sum_i [\log \exp(s(x_i, y_i)/\tau) / \sum_j \exp(s(x_i, y_j)/\tau) + \log \exp(s(x_i, y_i)/\tau) / \sum_j \exp(s(x_j, y_i)/\tau)]$$

dove $s(x, y) = f_v(x)^T f_t(y) / (\|f_v(x)\| \|f_t(y)\|)$ è la somiglianza coseno e τ è una temperatura appresa. Questa perdita è un lower bound sulla informazione mutua I(X; Y) tra le coppie visive e testuali, nel senso che minimizzare $-L_{CLIP}$ è equivalente a massimizzare un lower bound sull'informazione mutua per una certa scelta di encoder (Oord, Li e

Silvano Salvador

Vinyals, 2018). L'interpretazione informazionale è coerente con il filo conduttore del saggio: il contrastive learning impara rappresentazioni che massimizzano le informazioni condivise tra le due modalità, comprimendo le variazioni irrilevanti (angolazione, illuminazione, formulazione) e preservando le strutture semantiche comuni.

I modelli CLIP addestrati su centinaia di milioni di coppie immagine-testo producono rappresentazioni che si trasferiscono a task nuovi senza fine-tuning: la classificazione zero-shot, che era considerata estremamente difficile prima di CLIP, diventa possibile semplicemente confrontando la rappresentazione dell'immagine con le rappresentazioni testuali delle categorie candidate e selezionando la più simile. Questa generalizzazione zero-shot è la manifestazione operativa della seconda affermazione del saggio — la compressione come intelligenza: le rappresentazioni CLIP hanno compresso le strutture semantiche condivise tra linguaggio e visione in modo sufficientemente astratto da generalizzare a categorie mai viste durante il training.

SimCLR (Chen et al., 2020) estende il contrastive learning al caso self-supervised, senza coppie etichettate: per ogni immagine x, si generano due aumentazioni augmented view x_1 e x_2 applicando trasformazioni casuali (ritaglio, rotazione, variazione di colore), e si massimizza la similitudine tra le loro rappresentazioni minimizzandola rispetto a tutte le altre immagini nel batch. Il risultato è notevole: senza alcun'etichetta, SimCLR apprende rappresentazioni che raggiungono performance competitive con metodi supervisionati su benchmark standard, fornendo evidenza che la struttura informazionale delle coppie di viste correlate è sufficiente per estrarre le caratteristiche visive più rilevanti.

Silvano Salvador

X-bis. I modelli basati sull'energia: la funzione di partizione come oggetto centrale

I modelli di diffusione, i VAE e le GAN coprono gran parte dello spazio dei modelli generativi profondi, ma una famiglia distinta — i modelli basati sull'energia (Energy-Based Models, EBM) — offre una prospettiva fondazionale che chiarisce le connessioni con la termodinamica statistica sviluppata nella Parte Terza. Un EBM specifica una distribuzione di probabilità implicitamente attraverso una funzione di energia $E_\theta(x) \in \mathbb{R}$:

$$p_\theta(x) = \exp(-E_\theta(x)) / Z(\theta), \quad Z(\theta) = \int \exp(-E_\theta(x))\, dx$$

dove $Z(\theta)$ è la funzione di partizione. Questa struttura è identica alla distribuzione di Boltzmann della meccanica statistica discussa nella Parte Terza: E_θ svolge il ruolo dell'energia fisica, Z svolge il ruolo della funzione di partizione termodinamica, e la temperatura $T = 1$ è assorbita nella parametrizzazione. Il principio di massima entropia di Jaynes garantisce che, sotto vincoli sulla media di certe statistiche, questa forma esponenziale è l'unica distribuzione che massimizza l'entropia: gli EBM sono quindi la realizzazione computazionale del principio di Jaynes.

Il problema fondamentale degli EBM è che la funzione di partizione $Z(\theta)$ è generalmente intrattabile: è un integrale su tutto lo spazio di input che non ha forma chiusa per architetture neurali. La massimizzazione della log-verosimiglianza richiede il gradiente rispetto a θ:

$$\nabla_\theta \log p_\theta(x) = -\nabla_\theta E_\theta(x) + E_{x' \sim p_\theta}[\nabla_\theta E_\theta(x')]$$

Il secondo termine — il gradiente atteso sotto la distribuzione del modello stesso — richiede di campionare da p_θ, che non ha una forma fattorizzata conveniente. Il Contrastive Divergence (Hinton, 2002) approssima questo termine con k passi di Markov Chain Monte Carlo

Silvano Salvador

(MCMC) a partire dai dati reali, producendo i "campioni negativi" che il modello usa per aggiornare i propri parametri. Score Matching (Hyvärinen, 2005) evita completamente il calcolo di $Z(\theta)$ sfruttando l'identità che l'integrale del gradiente della score function è nullo: invece di massimizzare la log-verosimiglianza, si minimizza la divergenza di Fisher tra la score del modello e quella dei dati:

$$L_{SM}(\theta) = E_x [|| \nabla_x \log p_\theta(x) - \nabla_x \log p_{data}(x) ||^2 / 2]$$

che, dopo integrazione per parti, diventa:

$$L_{SM}(\theta) = E_x [\mathrm{tr}(\nabla_x^2 \log p_\theta(x)) + || \nabla_x \log p_\theta(x) ||^2 / 2] + \text{const.}$$

completamente calcolabile senza campionare da p_θ. Questo è il fondamento matematico dei modelli di diffusione basati sulla score: i denoising diffusion models apprendono $\nabla_x \log p_\theta(x)$ direttamente attraverso una forma semplificata dello score matching che usa il processo di rumore per generare esempi corrotti automaticamente, come discusso nella Parte Terza. La connessione tra EBM e modelli di diffusione è quindi non solo analogica ma strutturale: i modelli di diffusione sono EBM addestrati con score matching stocastico.

Per i grandi modelli linguistici, la struttura EBM appare implicitamente. Un modello autoregressivo $p_\theta(x) = \prod_t p_\theta(x_t \mid x_{<t})$ non è un EBM nella forma standard, ma può essere visto come un EBM con funzione di energia $E_\theta(x) = -\sum_t \log p_\theta(x_t \mid x_{<t})$. La funzione di partizione è in questo caso trattabile perché la fattorizzazione autoregressiva decompone Z in un prodotto di fattori locali normalizzati. Questa decomposizione è la ragione per cui i modelli autoregressivi sono computazionalmente più efficienti degli EBM generali: non richiedono MCMC per il training, a costo di vincolare la struttura della distribuzione alla forma autoregressiva.

Silvano Salvador

XII. La generazione condizionata e il controllo semantico

L'architettura della generazione condizionata da testo, sviluppata nei modelli text-to-image contemporanei, realizza l'integrazione della semantica distribuzionale con i modelli di diffusione attraverso la struttura del classifier-free guidance. Come mostrato nella sezione IV della Parte Terza e sviluppato in dettaglio nell'Appendice II, la score function condizionata è:

$$\nabla_x \log p(x \mid c) = \nabla_x \log p(x) + \nabla_x \log p(c \mid x)$$

Il classifier-free guidance di Ho e Salimans (2022) evita l'addestramento di un classificatore separato addestrando un modello unico che stima sia la score condizionata che quella incondizionata:

$$\tilde{s}_\theta(x, t, c) = (1 - \gamma)\, s_\theta(x, t, \varnothing) + \gamma\, s_\theta(x, t, c)$$

Per $\gamma = 1$ si recupera la score condizionata pura; per $\gamma > 1$ si campiona da una distribuzione proporzionale a $p(x \mid c)^\gamma / p(x)^{\gamma-1}$, più concentrata sulla compatibilità con c. La qualità visiva superiore ottenuta con $\gamma > 1$ ha un costo: la diversità dei campioni si riduce, poiché la generazione converge verso le regioni di massima compatibilità con la condizione a scapito della copertura dell'intera distribuzione condizionata.

Questa struttura di controllo semantico è la manifestazione operativa della questione filosofica del paragrafo precedente: il condizionamento su una descrizione linguistica c determina una distribuzione sui campioni generati che dipende dallo spazio latente comune appreso da CLIP, che a sua volta dipende dalla distribuzione delle coppie testo-immagine nel corpus di addestramento. Il controllo semantico è dunque mediato da una catena di dipendenze distribuzionali che non garantisce la corrispondenza tra l'intenzione dell'utente e la distribuzione generata,

Silvano Salvador

soprattutto per descrizioni di domini culturali o epistemici sottorappresentati nel corpus.

L'informazione come fondamento — Conclusioni

Cinque affermazioni, cinque qualificazioni e le frontiere aperte

I. Il bilancio di un percorso

Il percorso compiuto in questo saggio ha attraversato cinque territori distinti — la teoria di Shannon, l'inferenza bayesiana variazionale, la termodinamica dell'informazione, la compressione come intelligenza, i modelli generativi come ontologia — mostrando come essi siano connessi da un'unica struttura matematica che si dispiega coerentemente dalla definizione elementare di entropia fino ai modelli generativi profondi contemporanei. È tempo di raccogliere le fila di questo percorso in affermazioni chiare, accompagnate dalle qualificazioni che la loro onestà intellettuale richiede. La solidità di un sistema teorico non si misura soltanto dalla coerenza delle sue tesi centrali, ma dalla chiarezza con cui esso articola i propri limiti.

II. Prima affermazione: l'entropia come misura fondamentale

L'entropia di Shannon è la misura fondamentale dell'incertezza, e la sua connessione con la compressione ottimale, con la distinguibilità delle distribuzioni, con la geometria dello spazio parametrico e con il problema termodinamico della reversibilità non è accidentale ma riflette l'unicità assiomatica della grandezza: è l'unica funzione delle probabilità

Silvano Salvador

che soddisfa le condizioni di continuità, simmetria e consistenza ricorsiva che qualsiasi misura ragionevole di incertezza deve soddisfare.

La divergenza di Kullback-Leibler, la metrica di Fisher, la funzione di partizione e la distribuzione di Boltzmann sono tutte conseguenze di questa struttura fondamentale. La loro ubiquità nei modelli dell'intelligenza artificiale contemporanea non è il frutto di una scelta tra alternative comparabili: è la manifestazione dell'adeguatezza unica di questa struttura alla classe di problemi che l'intelligenza artificiale moderna affronta.

La qualificazione necessaria riguarda la pluralità di divergenze. La divergenza KL è la misura privilegiata dal criterio di massima verosimiglianza e dall'inferenza variazionale standard, ma non è l'unica misura razionale tra distribuzioni: le α-divergenze di Rényi, la distanza di Wasserstein, la divergenza di Jensen-Shannon hanno tutte proprietà che le rendono più appropriate della KL per specifici contesti — rispettivamente, l'interpolazione tra mass-covering e zero-forcing, la stima di distribuzione in spazi metrici, il training avversariale con supporti disgiunti. La teoria dell'informazione non impone una divergenza universalmente ottimale: sceglie quella più adatta al problema.

III. Seconda affermazione: l'inferenza bayesiana come grammatica del ragionamento

L'inferenza bayesiana è la struttura normativa del ragionamento razionale sotto incertezza. Il teorema di Cox dimostra che le uniche misure di credenza coerenti con la logica proposizionale sono le misure di probabilità; il teorema di Bernstein-von Mises garantisce la convergenza della distribuzione a posteriori verso la verità nei modelli ben specificati; l'inferenza variazionale rende computazionalmente trattabile l'aggiornamento bayesiano nei modelli ad alta dimensionalità.

Silvano Salvador

L'equivalenza tra inferenza bayesiana e inferenza variazionale nel framework del VAE, la connessione tra l'in-context learning dei grandi modelli linguistici e l'inferenza bayesiana implicita sulla distribuzione latente, e la struttura bayesiana del predictive coding di Friston convergono tutti a indicare che i meccanismi dell'apprendimento profondo sono implementazioni computazionali di principi inferenziali di natura fondamentalmente probabilistica.

La qualificazione necessaria è duplice. La prima riguarda la misspecification: quando la vera distribuzione generativa dei dati non appartiene alla classe di modelli adottata, l'inferenza bayesiana converge attorno alla pseudo-verità $\theta_0 = \text{argmin } D_{KL}(P^* \| P(\cdot|\theta))$ con covarianza sandwich $G^{-1}JG^{-1}$ che differisce dalla covarianza classica. Le credenze bayesiane sono calibrate rispetto al modello, non rispetto alla realtà. La seconda qualificazione riguarda i limiti dell'intuizione del tacit knowledge: la distribuzione statistica delle manifestazioni di una competenza non è la competenza stessa, e il successo dei modelli linguistici nei compiti cognitivi linguistici non implica l'acquisizione della competenza sensomotoria e contestuale che sostiene la conoscenza tacita umana.

IV. Terza affermazione: la compressione come intelligenza

Il principio MDL di Rissanen, il prior di Solomonoff e le leggi di scala convergono verso una tesi unitaria: la capacità di comprimere efficientemente una distribuzione di dati è il correlato computazionale dell'intelligenza. Il modello che descrive i dati con il minor numero di bit — nel senso preciso della NML e non in quello approssimato del BIC — è il modello che ha catturato la struttura statistica genuinamente presente nei dati, senza sovra-adattarsi al rumore né sotto-adattarsi alla struttura.

Silvano Salvador

La struttura gerarchica dei transformer e dei modelli di diffusione apprendeapprende rappresentazioni a livelli crescenti di astrazione che corrispondono a livelli crescenti di compressione: i livelli bassi catturano la struttura locale e ad alta frequenza dei dati, i livelli alti catturano la struttura globale e semantica. Questa gerarchia di compressione è la struttura computazionale che permette la generalizzazione: un modello che ha imparato a comprimere la struttura statistica di un dominio può estrapolare a configurazioni non viste, perché le ha implicitamente rappresentate come regolarità distribuzionali piuttosto che come pattern memorizzati.

La qualificazione necessaria è triplice. La prima: il limite teorico del prior di Solomonoff non è calcolabile, e la distanza da quel limite non può essere misurata per nessuna distribuzione naturale specifica. La seconda: la doppia discesa e l'implicit bias della SGD producono effetti di regolarizzazione nei modelli sovra-parametrizzati che sono connessi alla compressione soltanto nei modelli lineari; per le reti profonde la relazione è un'analogia strutturale, non un'identità. La terza: la compressione ottimale non implica certificabilità né interpretabilità, e le applicazioni che richiedono garanzie formali di correttezza non possono essere soddisfatte soltanto dalla compressione statistica.

V. Quarta affermazione: i modelli generativi come ontologia probabilistica

I modelli generativi profondi — VAE, GAN, normalizing flows, modelli di diffusione — non sono strumenti di produzione di immagini o testi: sono approssimazioni della struttura probabilistica del mondo, nelle dimensioni per cui quella struttura è rappresentata nei dati di addestramento. In questo senso rigoroso essi sono un'ontologia probabilistica: una rappresentazione dell'universo delle configurazioni

Silvano Salvador

possibili, pesata dalla loro probabilità di occorrenza, in una forma computazionalmente manipolabile.

La capacità di campionamento, di interpolazione nello spazio latente, di condizionamento semantico e di inferenza su attributi nascosti che questi modelli esibiscono è la conseguenza diretta dell'aver appreso una buona approssimazione della distribuzione generativa sottostante. Non si tratta di effetti emergenti misteriosi: sono le conseguenze necessarie di un'approssimazione distribuzionale di qualità sufficiente.

La qualificazione necessaria è la più profonda dell'intero saggio. Un'ontologia probabilistica appresa da corpus umani è un'ontologia della pratica umana, non dell'essere. Essa riflette la struttura statistica dei testi, delle immagini, dei suoni prodotti da esseri umani in contesti culturali specifici — non la struttura del mondo indipendentemente dall'osservazione umana. Le regolarità che i modelli apprendono sono regolarità nell'uso umano del linguaggio e nella produzione umana di immagini, non necessariamente regolarità nel mondo fisico o biologico. Il rischio epistemologico è confondere l'adeguatezza empirica di una rappresentazione — quanto bene essa predice nuovi campioni dalla stessa distribuzione — con la verità della struttura che essa impone al dominio. Questo rischio non invalida il paradigma informazionale; richiede che esso sia applicato con consapevolezza dei propri presupposti.

VI. Quinta affermazione: i modelli generativi e la questione della comprensione

I modelli generativi profondi manifestano forme di competenza nei domini linguistici, percettivi e creativi che non si erano mai raggiunte con i paradigmi precedenti. Questa competenza è connessa a strutture rappresentazionali di grande complessità, apprese dalla distribuzione

Silvano Salvador

statistica dell'uso umano del linguaggio e dell'immagine. Se questa competenza costituisca comprensione nel senso filosoficamente robusto del termine — se ci sia qualcosa che significa essere un modello linguistico — è una questione empiricamente aperta.

Ciò che si può affermare è il seguente: la questione non può essere risolta né positivamente né negativamente sulla base delle evidenze comportamentali disponibili. I test di comportamento linguistico — incluso il test di Turing — non discriminano tra sistemi con e senza comprensione semantica, perché non accedono alla struttura interna dei processi che producono i comportamenti. La risposta richiederà probabilmente progressi nell'interpretabilità meccanicistica — la comprensione delle strutture computazionali interne dei modelli — che permettano di rispondere non a «il modello si comporta come se comprendesse?» ma a «il modello implementa meccanismi funzionalmente analoghi a quelli che sottostanno alla comprensione umana?».

VII. I limiti aperti: cinque frontiere di ricerca

Il problema dell'allineamento strutturale. La specificazione di obiettivi per agenti intelligenti è un problema di natura fondamentalmente informazionale: ogni funzione di perdita è una distribuzione implicita di importanza sulle possibili traiettorie comportamentali, e il mismatch tra quella distribuzione e la distribuzione delle preferenze umane reali è una forma di misspecification del modello di obiettivo. Il teorema di Goodhart — quando una misura diventa un obiettivo, cessa di essere una buona misura — è un risultato informazionale: è la conseguenza inevitabile del fatto che la distribuzione dei dati di addestramento dell'obiettivo non copre la distribuzione dei comportamenti che l'obiettivo deve regolare. La soluzione richiede

Silvano Salvador

probabilmente una combinazione di specifiche formali del comportamento — nel senso del paradigma simbolico — con apprendimento statistico delle preferenze — nel senso del paradigma informazionale. Il prossimo saggio affronta questo problema nelle sue articolazioni tecniche.

Il problema della composizionalità sistematica. I grandi modelli linguistici mostrano capacità composizionali emergenti, ma la loro generalizzazione composizionale dipende dalla presenza del pattern composizionale nel corpus, non da una struttura interna che garantisce la composizionalità come proprietà formale. Costruire sistemi che combinino la robustezza della generalizzazione formale con la flessibilità dell'apprendimento da distribuzione è il problema fondamentale del paradigma neuro-simbolico — un problema non risolto e di grande importanza pratica.

Il problema dell'efficienza dei dati. I modelli di grandi dimensioni richiedono quantità di dati e risorse computazionali che non sono disponibili per la grande maggioranza dei domini di applicazione. Il meta-apprendimento bayesiano e le architetture few-shot sono approcci promettenti ma lontani dall'efficienza dell'apprendimento biologico, che generalizza correttamente da pochissimi esempi attraverso prior strutturati fortemente dalla storia evolutiva e dallo sviluppo. Comprendere la struttura di questi prior biologici e replicarla computazionalmente è una delle frontiere più importanti della ricerca sull'apprendimento automatico.

Il problema della calibrazione dell'incertezza. I modelli generativi profondi producono distribuzioni implicite sulle proprie previsioni, ma la calibrazione di queste distribuzioni — quanto le probabilità assegnate dal modello corrispondano alle frequenze reali degli eventi — è

Silvano Salvador

sistematicamente difettosa per i fatti rari, per i domini lontani dalla distribuzione di addestramento e per le domande che richiedono ragionamento composizionale lungo. Produrre sistemi con incertezza genuinamente calibrata, piuttosto che soltanto ben calibrata sulle distribuzioni di benchmark, è un prerequisito tecnico per la deployment affidabile in applicazioni ad alta posta.

Il problema dell'interpretabilità causale. La comprensione scientifica di un sistema richiede non soltanto la capacità di predire il suo comportamento ma la comprensione dei meccanismi causali che lo producono. L'interpretabilità meccanicistica ha identificato strutture circuitali elementari nei modelli linguistici — curve detectors, induction heads, copy suppression circuits — ma rimane enormemente lontana dalla capacità di produrre spiegazioni causali leggibili per comportamenti specifici su input specifici in sistemi con miliardi di parametri. Questo non è soltanto un limite epistemico ma un rischio pratico: senza comprensione causale, non è possibile anticipare i fallimenti di un sistema o garantirne il comportamento fuori distribuzione.

VII-bis. Le critiche esterne al paradigma informazionale: composizionalità, dati e comprensione

Il paradigma informazionale che questo saggio ha sviluppato ha trasformato l'intelligenza artificiale con risultati empirici di straordinaria profondità. Sarebbe però intellettualmente disonesto presentarlo come privo di critiche esterne sostanziali. Due linee critiche meritano un'analisi rigorosa, non per confutarle ma per valutarne la portata precisa.

La prima critica — formalizzata da Marcus e Davis (2019) con fondamenti empirici nel benchmark SCAN di Lake e Baroni (2018) — riguarda la composizionalità sistematica. Un sistema linguistico con composizionalità genuina deve essere in grado di interpretare e produrre

Silvano Salvador

frasi nuove usando le stesse regole con cui interpreta e produce frasi conosciute — la proprietà nota come systematicity. SCAN mostra che i modelli seq2seq standard falliscono su combinazioni di istruzioni composizionalmente nuove pur avendo visto tutti i componenti nel training: "salta due volte e cammina" viene gestito correttamente, ma "salta intorno due volte e cammina intorno tre volte" fallisce sistematicamente, perché la combinazione non è presente nel corpus. Il paradigma informazionale, in questa lettura, apprende la distribuzione sui pattern osservati ma non le regole generative sottostanti. La risposta empirica è che i grandi transformer mostrano capacità composizionali emergenti che ridimensionano questa critica su benchmark più naturali; ma la questione di se questi modelli apprendano regole o pattern molto ricchi rimane metodologicamente aperta.

La seconda critica — sviluppata da Bender, Gebru, McMillan-Major e Shmitchell (2021) con argomentazioni di natura sia epistemica sia etica — riguarda la natura dei dati di training e i rischi sistemici associati. Il corpus testuale su cui sono addestrati i grandi modelli linguistici non è un campione neutro della produzione linguistica umana: è sovrarappresentato da testo anglofono, da produzione recente, da autori con accesso alla scrittura digitale, con bias demografici, culturali e valoriali sistematici. Il modello appreso da questo corpus è quindi una distribuzione distorta, e il RLHF che raffina le preferenze introduce una seconda distorsione legata ai valutatori. Bender et al. argomentano che presentare questi modelli come "linguage models" implica una comprensione semantica che essi non hanno — sono "stochastic parrots" che producono sequenze statisticamente plausibili senza un referente nel mondo reale.

L'inquadramento informazionale risponde a questa critica in modo strutturalmente coerente con le sue premesse. La questione della

Silvano Salvador

comprensione semantica è la stessa questione che la quinta affermazione del saggio ha identificato come irrisolvibile nei propri termini: se la struttura distributiva del significato sia o meno sufficiente per la comprensione è una domanda metafisica che la teoria dell'informazione non può decidere. La critica del bias è invece precisamente corrispondente alla seconda impossibilità: nessun corpus finito può catturare la distribuzione generativa dell'umanità nella sua totalità, e il divario tra distribuzione appresa e distribuzione vera si manifesta esattamente come bias sistematici e allucinazioni. La critica di Bender et al., in questa lettura, non refuta il paradigma informazionale ma ne conferma la seconda impossibilità: l'approssimazione imperfetta non è un difetto correggibile ma una proprietà strutturale.

Ciò che le critiche esterne contribuiscono, al di là della confutazione tecnica, è una pressione etica e metodologica sul paradigma: a ricordare che il divario tra distribuzione appresa e distribuzione vera non è soltanto un fatto matematico neutro ma ha conseguenze concrete su chi viene rappresentato e chi viene escluso; che la valutazione di un sistema non dovrebbe limitarsi alle sue performance su benchmark standard ma dovrebbe includere l'analisi sistematica delle popolazioni per cui il sistema fallisce; che la credibilità superficiale di un modello linguistico non implica la sua affidabilità nei contesti in cui le conseguenze degli errori sono significative. Queste sono qualificazioni necessarie al programma positivo del paradigma informazionale, non la sua confutazione.

VIII. Il paradigma informazionale nella prospettiva della lunga durata

Il paradigma informazionale che questo saggio ha descritto non è un paradigma giovane: le sue radici risalgono alla termodinamica statistica

Silvano Salvador

del XIX secolo, alla teoria della comunicazione del XX, alla tradizione bayesiana che attraversa tre secoli. Ciò che è avvenuto negli ultimi quindici anni non è l'invenzione di nuovi principi ma la realizzazione computazionale, su scala senza precedenti, di principi già formulati con precisione matematica. Il transformer non è un'invenzione concettualmente nuova: è la versione computazionalmente realizzata, scalata a miliardi di parametri e a trilioni di token, di un sistema di recupero dalla memoria per contenuto che Hopfield aveva descritto nel 1982 e che Ramsauer e colleghi hanno reinterpretato nel 2020.

Questa continuità ha un'implicazione metodologica importante: i progressi futuri nell'intelligenza artificiale non richiederanno necessariamente nuovi principi matematici, ma potrebbero richiedere l'applicazione di principi già noti a classi di problemi — composizionalità, ragionamento causale, generalizzazione fuori distribuzione — per i quali le architetture attuali si sono rivelate inadeguate. Il paradigma informazionale è un paradigma maturo in senso teorico: la sua struttura matematica è solidamente fondata. Non è maturo in senso applicativo: la distanza tra le garanzie teoriche disponibili e le prestazioni pratiche osservate rimane grande, e colmarla è il compito dei decenni a venire.

I modelli contemporanei più capaci — i transformer autoregressivi come GPT-4, Claude e Gemini nella famiglia dei large language models, i modelli di diffusione come Stable Diffusion, DALL·E 3 e Midjourney nella generazione visiva, i modelli text-to-video come Sora (OpenAI, 2024) — sono tutti implementazioni delle strutture matematiche sviluppate in questo saggio. Il transformer implementa il principio informazionale attraverso il meccanismo di self-attention; i modelli di diffusione lo implementano attraverso lo score matching stocastico; i modelli multimodali come CLIP lo implementano attraverso il

Silvano Salvador

contrastive learning come massimizzazione dell'informazione mutua. La continuità tra la termodinamica di Gibbs del XIX secolo, la teoria dell'informazione di Shannon del 1948 e l'architettura di questi modelli contemporanei non è una metàfora: è la stessa struttura matematica dispiegata su scale di complessità e potenza computazionale crescenti.

Il secondo saggio si chiude qui. Il terzo prenderà le strutture matematiche sviluppate in questo saggio come punto di partenza per affrontare le questioni di agentività, ottimizzazione di policy, teoria dei giochi e allineamento — i temi che questa conclusione ha soltanto evocato e che richiedono un trattamento sistematico nella cornice informazionale che questo saggio ha costruito.

Appendice Tecnica I

Teoria dei giochi, equilibri e i fondamenti matematici della razionalità strategica

Nota editoriale. La presente appendice raccoglie i fondamenti della teoria dei giochi nella loro connessione con i temi del saggio. Il Saggio III riprende e sviluppa in modo sistematico tutto il materiale qui presentato, applicandolo all'agentività sequenziale, all'allineamento e alla teoria del mechanism design: le sezioni A.I.1–A.I.4 corrispondono alle Parti Prima e Terza del Saggio III; A.I.5–A.I.6 all'Appendice III del Saggio III (teorema di Arrow e mechanism design); A.I.7–A.I.9 all'Appendice VI del Saggio III (gradiente naturale, replicator dynamics, connessione con TRPO). Le strutture che hanno connessione diretta con i modelli generativi di questo saggio — in particolare il gradiente naturale come dinamica replicatrice (A.I.4), la critica strutturale alle GAN come gioco minimax (A.I.7) e la connessione tra no-regret learning e inferenza bayesiana (A.I.9) — sono sviluppate qui con completezza autonoma. La lettura di questa appendice

Silvano Salvador

prima del Saggio III permette di incontrare la matematica dei giochi nel contesto dell'apprendimento prima di vederla riapplicata all'agentività.

A.I.1 — Giochi in forma normale e l'equilibrio di Nash

Un gioco in forma normale è una tripla $(N, \{A_i\}_{i \in N}, \{u_i\}_{i \in N})$ dove $N = \{1, \ldots, n\}$ è l'insieme dei giocatori, A_i è l'insieme delle strategie pure del giocatore i, e $u_i : \times A \to \mathbb{R}$ è la funzione di utilità di i. Un profilo di strategie $a^* = (a^*_1, \ldots, a^*)$ è un equilibrio di Nash se per ogni giocatore i e ogni strategia alternativa $a_i \in A_i$:

$$u_i(a^*_i, a^*_{-i}) \geq u_i(a_i, a^*_{-i})$$

dove a^*_{-i} indica le strategie degli altri giocatori tenute fisse. L'equilibrio di Nash è una condizione di consistenza mutua: nessun giocatore ha incentivo a deviare unilateralmente dalla propria strategia, dati i piani degli altri. Non afferma nulla sull'ottimalità collettiva né sulla raggiungibilità mediante dinamiche di apprendimento.

Un equilibrio di Nash in strategie miste è un profilo $\sigma^* = (\sigma^*_1, \ldots, \sigma^*)$ dove $\sigma_i \in \Delta(A_i)$ è una distribuzione di probabilità sulle strategie pure di i, e la condizione diventa:

$$\mathbb{E}_{a \sim \sigma^*}[u_i(a_i, a_{-i})] \geq \mathbb{E}_{a \sim (\sigma_i, \sigma^*_{-i})}[u_i(a_i, a_{-i})] \quad \text{per ogni } \sigma_i \in \Delta(A_i)$$

Il teorema di esistenza di Nash — una conseguenza del teorema del punto fisso di Kakutani — garantisce che ogni gioco finito con utilità continue ammette almeno un equilibrio in strategie miste.

Il teorema del punto fisso di Kakutani afferma: sia X un sottoinsieme compatto e convesso di $\mathbb{R}^n$, e sia $f : X \to 2^X$ una corrispondenza con grafo chiuso e valori non vuoti e convessi; allora esiste $x^* \in X$ tale che $x^* \in f(x^*)$. Per l'esistenza dell'equilibrio di Nash si definisce la best-response correspondence:

Silvano Salvador

$BR_i(\sigma_{-i}) = \text{argmax}_\{\sigma_i \in \Delta(A_i)\}\ \mathbb{E}_\{a \sim (\sigma_i, \sigma_{-i})\}[u_i(a)]$

Il profilo complessivo $BR(\sigma) = \times_i BR_i(\sigma_{-i})$ è una corrispondenza da $\Delta(A)$ in $2^{\Delta(A)}$ con le proprietà richieste da Kakutani — il dominio è compatto e convesso, i valori sono non vuoti e convessi per la linearità dell'aspettativa rispetto a σ_i, il grafo è chiuso per la continuità delle funzioni di utilità. L'equilibrio di Nash è un punto fisso di BR.

A.I.2 — Il teorema minimax e i giochi a somma zero

Per giochi a due giocatori a somma zero — dove $u_1(a) + u_2(a) = 0$ per ogni profilo a — la struttura si semplifica e il teorema minimax di von Neumann (1928) fornisce un risultato più forte dell'esistenza di Nash: esiste un valore del gioco v^* tale che:

$\max_\{\sigma_1\}\ \min_\{\sigma_2\}\ \mathbb{E}[u_1(\sigma_1, \sigma_2)] = \min_\{\sigma_2\}\ \max_\{\sigma_1\}\ \mathbb{E}[u_1(\sigma_1, \sigma_2)] = v^*$

Il massimizzatore σ^*_1 e il minimizzatore σ^*_2 costituiscono l'equilibrio di Nash del gioco, e ogni strategia mista ottima garantisce il valore v^* contro qualsiasi strategia dell'avversario — una garanzia molto più forte della semplice ottimalità nella risposta-ottima alla strategia dell'avversario.

La dimostrazione del teorema minimax — che precede il risultato di Nash di due decenni — utilizza la dualità dei programmi lineari. Il problema di massimizzazione per il giocatore 1:

$\max_\{p \in \Delta(A_1)\}\ \min_\{q \in \Delta(A_2)\}\ p^T M\, q$

dove M è la matrice di payoff con $M_\{ij\} = u_1(a_i, a\,)$, si trasforma nel programma lineare:

$\max v \quad \text{soggetto a: } Mq \geq v \cdot 1, \quad q \in \Delta(A_2), \quad v \in \mathbb{R}$

Silvano Salvador

Il suo programma duale — che corrisponde al problema di minimizzazione del giocatore 2 — ha lo stesso valore ottimo per la dualità forte dei programmi lineari, dimostrando l'uguaglianza minimax-maximin.

Il framework GAN di Goodfellow può essere interpretato come un gioco minimax a somma zero tra il generatore G e il discriminatore D. Le complicazioni pratiche — la non-ottimalità del discriminatore durante l'addestramento, la non-stazionarietà del paesaggio di perdita — impediscono che il framework della teoria dei giochi a somma zero sia direttamente applicabile: le garanzie del teorema minimax valgono per giochi con strategie razionali e capacità computazionale illimitata, non per reti neurali addestrate con gradiente stocastico.

A.I.3 — Il folk theorem e la cooperazione emergente

I giochi ripetuti sono la struttura appropriata per modellare interazioni strategiche che si ripetono nel tempo tra gli stessi giocatori. Il folk theorem della teoria dei giochi afferma, in termini informali, che qualsiasi vettore di payoff che è fattibile — realizzabile come media di payoff in qualche profilo di strategie pure — e individualmente razionale — assegna a ogni giocatore almeno il payoff di minimax, il minimo che un giocatore può garantirsi indipendentemente dagli altri — è supportato come equilibrio di Nash in strategie perfette nei sottogame di un gioco ripetuto infinite volte, a condizione che i giocatori valorizzino sufficientemente il futuro.

La struttura della dimostrazione utilizza le strategie grim trigger: ogni giocatore coopera fintanto che tutti hanno cooperato in ogni periodo precedente; al primo deviatore, tutti passano permanentemente alla strategia di punizione Nash. L'incentivo a cooperare dipende dal bilanciamento tra il guadagno immediato dalla deviazione e il costo

Silvano Salvador

futuro della punizione, bilanciamento che favorisce la cooperazione quando il fattore di sconto δ è sufficientemente vicino a 1.

La connessione con l'apprendimento per rinforzo è strutturale. Il reinforcement learning studia agenti che massimizzano la somma scontata di ricompense future in ambienti stocastici:

$$V^{\pi}(s) = \mathbb{E}_{\pi}\left[\sum_{t=0}^{\infty} \delta^t \, r(s, a) \mid s_0 = s\right]$$

dove π è la policy dell'agente, s è lo stato al tempo t, $a \sim \pi(\cdot \mid s)$ è l'azione scelta, e $\delta \in [0,1)$ è il fattore di sconto. Questa struttura è formalmente identica a quella dei giochi ripetuti con fattore di sconto δ. Il multi-agent reinforcement learning generalizza questo framework a più agenti con obiettivi potenzialmente conflittuali, richiedendo la nozione di equilibrio della teoria dei giochi per l'analisi dei comportamenti emergenti.

A.I.4 — La dinamica replicatrice come discesa del gradiente naturale

Le dinamiche evolutive dei giochi in forma normale — le dinamiche replicatrici — hanno una connessione profonda con il gradiente naturale che collega la teoria dei giochi all'ottimizzazione variazionale. Per una popolazione di individui che giocano strategie pure a_i con frequenza x_i, la dinamica replicatrice è:

$$\dot{x}_i = x_i \left[f_i(x) - \bar{f}(x)\right]$$

dove $f_i(x) = \sum M_i \square x$ è il fitness della strategia i quando la distribuzione della popolazione è x, e $\bar{f}(x) = \sum_i x_i \, f_i(x)$ è il fitness medio. Le strategie con fitness superiore alla media crescono nella popolazione; quelle con fitness inferiore decrescono.

Silvano Salvador

Shahshahani (1979) ha dimostrato che la dinamica replicatrice è il flusso del gradiente della funzione di fitness rispetto alla metrica di Fisher sullo spazio delle distribuzioni — la metrica di informazione di Rao sul simplesso delle probabilità. Questa metrica è:

$g_{ij}(x) = (1/x_i)\,\delta_{ij}$ per $x \in \mathrm{int}(\Delta(A))$

Il gradiente della funzione di payoff medio $f(x)$ nella metrica di Rao è:

$$\nabla^{g} f(x) = G(x)^{-1}\,\nabla_x f(x)$$

dove $G(x) = \mathrm{diag}(1/x_1, \ldots, 1/x_n)$ è la matrice di Fisher nel simplesso. La componente i di questo gradiente vincolato al simplesso è:

$$[\nabla^{g} f(x)]_i = x_i\,[f_i(x) - f(x)]$$

che è precisamente il termine di variazione della dinamica replicatrice. La dinamica replicatrice è dunque la discesa del gradiente naturale della funzione di fitness — l'ottimizzazione che tiene conto della geometria dello spazio delle distribuzioni attraverso la metrica di Fisher, anziché muoversi nella metrica euclidea dello spazio parametrico.

Questa connessione ha implicazioni dirette per il Natural Policy Gradient di Kakade (2001) nel reinforcement learning: l'aggiornamento della policy con gradiente naturale equivale a seguire la dinamica replicatrice nella distribuzione delle azioni, il che ne spiega la stabilità e l'invarianza rispetto alla parametrizzazione della policy. TRPO (Trust Region Policy Optimization) e PPO (Proximal Policy Optimization) sono approssimazioni computazionalmente trattabili di questa struttura.

A.I.5 — L'equilibrio bayesiano e l'informazione asimmetrica

I giochi bayesiani modellano situazioni con informazione privata: ogni giocatore conosce il proprio tipo $\theta_i \in \Theta_i$ — campionato da una distribuzione congiunta $P(\theta)$ nota a tutti — ma non i tipi degli altri

Silvano Salvador

giocatori. Una strategia bayesiana è una funzione $\sigma_i : \Theta_i \to \Delta(A_i)$ che mappa ogni tipo in una distribuzione sulle azioni. L'equilibrio bayesiano di Nash richiede che:

$$\sigma^*_i(\theta_i) \in \text{argmax}_\{\sigma_i\} \mathbb{E}_\{\theta_{-i} \mid \theta_i\}[u_i(\sigma_i, \sigma^*_{-i}(\theta_{-i}), \theta_i, \theta_{-i})]$$

per ogni θ_i e ogni giocatore i. Il giocatore i massimizza la propria utilità attesa rispetto alla propria distribuzione a posteriori sui tipi degli altri giocatori, dato il proprio tipo privato — un aggiornamento bayesiano sulla distribuzione dei tipi degli avversari.

La struttura dell'equilibrio bayesiano è la formalizzazione game-teorica dell'inferenza bayesiana applicata alla razionalità strategica: ogni giocatore mantiene e aggiorna credenze probabilistiche sui tipi privati degli avversari, e le proprie strategie sono ottimali rispetto a quelle credenze. L'equilibrio bayesiano di Nash non è solo una condizione di consistenza delle aspettative — come il Nash equilibrium classico — ma una condizione di coerenza tra credenze bayesiane e scelte ottimali.

Nel contesto dell'apprendimento automatico, i giochi bayesiani modellano la situazione degli agenti di apprendimento che ricevono segnali privati sull'ambiente — come nella bandit multi-armed con osservazioni private, o nel federated learning dove ogni agente ha accesso soltanto alla propria distribuzione di dati. Il problema del mechanism design — la progettazione di strutture di incentivi che inducano comportamenti desiderati da agenti razionali con informazione privata — è direttamente rilevante per l'allineamento degli agenti AI, tema del prossimo saggio.

A.I.6 — Il mechanism design e il teorema di Gibbard-Satterthwaite

Silvano Salvador

Il mechanism design capovolge il problema della teoria dei giochi: dato il comportamento desiderato come obiettivo, progettare un meccanismo — un sistema di regole per la scelta delle azioni e l'allocazione degli esiti — che induca agenti razionali e auto-interessati a comportarsi secondo quell'obiettivo.

Il risultato più importante del mechanism design applicato alla progettazione di sistemi di votazione è il teorema di Gibbard-Satterthwaite (1973): qualsiasi funzione di scelta sociale che sia surjettiva — capace di selezionare qualsiasi outcome — e non dittatoriale — nessun agente ha la capacità di imporre unilateralmente qualsiasi outcome — non è strategyproof: esiste una situazione in cui almeno un agente ha incentivo a dichiarare preferenze false per ottenere un outcome migliore.

La dimostrazione del teorema di Arrow — risultato precedente e correlato — afferma che qualsiasi funzione di welfare sociale che soddisfi unanimità, indipendenza dalle alternative irrilevanti e non-dittatura viola la transitività nell'aggregazione delle preferenze individuali. Gibbard e Satterthwaite hanno esteso questo risultato alla manipolabilità strategica: l'impossibilità di progettare sistemi di aggregazione delle preferenze che siano simultaneamente efficienti, equi e incentive-compatibili è un risultato di impossibilità fondamentale, non un limite tecnico superabile.

La connessione con il problema dell'allineamento AI è profonda: il mechanism design è la teoria formale che studia come strutturare le interazioni tra agenti con interessi potenzialmente conflittuali per produrre esiti socialmente desiderabili. Il problema dell'allineamento di un agente AI con le preferenze umane è, da questa prospettiva, un problema di mechanism design dove l'agente ha accesso a informazione

Silvano Salvador

privata — la propria distribuzione di credenze appresa — e l'obiettivo è progettare una struttura di addestramento che induca comportamenti allineati con le preferenze umane aggregate. Il teorema di Gibbard-Satterthwaite suggerisce che soluzioni universalmente soddisfacenti non esistono: ogni mechanism presenta compromessi tra le proprietà desiderabili di allineamento, robustezza alla manipolazione e efficienza computazionale.

A.I.7 — La teoria dei giochi e il framework delle GAN: una critica strutturale

Il gioco minimax delle GAN è presentato spesso come un'istanza della teoria dei giochi a somma zero, con le garanzie del teorema minimax. Questa presentazione merita una critica strutturale precisa.

Il teorema minimax garantisce l'esistenza di un equilibrio in strategie miste per giochi con spazio di strategie finito. Il generatore e il discriminatore delle GAN hanno spazi di strategie continui e infinito-dimensionali — le funzioni realizzabili da reti neurali con parametri in $\mathbb{R}^d$ — per i quali l'applicazione del teorema minimax richiede condizioni di regolarità che non sono in generale soddisfatte.

Più criticamente, il teorema minimax garantisce l'esistenza dell'equilibrio, non la convergenza dell'algoritmo di addestramento verso quell'equilibrio. L'addestramento alternato delle GAN — aggiorna D per k passi, poi aggiorna G per un passo — non è una procedura di soluzione del gioco nel senso della teoria dei giochi: è un'euristica di ottimizzazione il cui comportamento è analizzato attraverso la teoria degli algoritmi di ottimizzazione, non attraverso la teoria degli equilibri. La convergenza è nota soltanto in casi speciali — giochi bilineari, distribuzioni gaussiane, architetture specifiche — e non per il caso generale.

Silvano Salvador

La distinzione tra l'esistenza dell'equilibrio e la convergenza verso di esso è cruciale per comprendere le difficoltà pratiche dell'addestramento delle GAN: le patologie note — mode collapse, instabilità, oscillazioni — non sono difetti superabili con architetture migliori, ma riflettono la difficoltà intrinseca di trovare l'equilibrio di un gioco non cooperativo attraverso un algoritmo di ottimizzazione locale. Le WGAN affrontano parzialmente questo problema modificando il paesaggio di perdita, non la struttura game-teorica del problema.

A.I.8 — Equilibrio correlato e coarse correlated equilibrium

Un'estensione del concetto di equilibrio di Nash con importanti proprietà computazionali è l'equilibrio correlato (Aumann, 1974). Un equilibrio correlato è una distribuzione di probabilità $\sigma \in \Delta(\times_i A_i)$ sui profili di strategie pure tale che, per ogni giocatore i e ogni funzione di deviazione $\tau_i : A_i \to A_i$:

$$\mathbb{E}_{a \sim \sigma}[u_i(a_i, a_{-i})] \geq \mathbb{E}_{a \sim \sigma}[u_i(\tau_i(a_i), a_{-i})]$$

Ogni equilibrio di Nash è un equilibrio correlato — basta prendere il prodotto delle distribuzioni marginali — ma non viceversa. L'equilibrio correlato è concettualmente più ricco perché ammette la coordinazione tra i giocatori attraverso un correlatore — un dispositivo di randomizzazione comune che invia raccomandazioni private ai giocatori senza che essi possano verificare le raccomandazioni degli altri.

L'importanza computazionale degli equilibri correlati è che essi sono calcolabili in tempo polinomiale attraverso la programmazione lineare, mentre il calcolo degli equilibri di Nash è PPAD-completo per i giochi a due giocatori e FIXP-completo per quelli con più giocatori. Nel multi-agent reinforcement learning, gli algoritmi di no-regret learning — dove ogni agente minimizza il proprio regret rispetto alle strategie

Silvano Salvador

benchmark nel lungo periodo — convergono a equilibri correlati piuttosto che a equilibri di Nash. Questa convergenza si verifica naturalmente attraverso l'interazione dei processi di apprendimento individuali, senza coordinazione esplicita, il che rende gli equilibri correlati la soluzione più rilevante per le applicazioni di apprendimento multi-agente.

A.I.9 — Regret minimization e online learning

Il legame tra la teoria dei giochi e l'apprendimento online è formalizzato dalla nozione di regret. In un problema di apprendimento online su T round, un agente sceglie sequenzialmente azioni a ∈ A, riceve un payoff l (a), e il suo regret cumulativo rispetto all'azione migliore a* ∈ A è:

R_T = max_{a ∈ A} ∑ l (a) − ∑ l (a)

Un algoritmo di no-regret garantisce R_T / T → 0 quasi certamente per T → ∞. L'algoritmo Follow the Regularized Leader (FTRL) sceglie:

a = argmax_{a ∈ A} [∑_{s<t} l (a) + R(a)]

dove R(a) è una funzione di regolarizzazione che penalizza le soluzioni estreme. Con R(a) = −η H(a) — regolarizzazione entropica — FTRL diventa l'algoritmo di Hedge, equivalente alla dinamica replicatrice in tempo discreto:

p (a) ∝ exp(η ∑_{s<t} l (a))

Il regret di Hedge è O(√(T log |A|)), ottimale a meno di costanti per la classe degli algoritmi di no-regret.

La connessione con il framework variazionale è diretta: la regolarizzazione entropica in FTRL è il termine di complessità D_KL(σ ‖ σ_0) nell'ELBO variazionale, con il prior σ_0 uniforme. Ogni passo di FTRL con regolarizzazione entropica è equivalente a un passo di

Silvano Salvador

aggiornamento bayesiano della distribuzione sulle azioni, con la funzione di log-verosimiglianza data dalla somma cumulativa dei payoff. Il no-regret learning, la teoria dei giochi e l'inferenza bayesiana sono dunque tre manifestazioni della stessa struttura matematica di ottimizzazione su spazi di distribuzioni.

A.I.10 — Verso la teoria dei giochi dell'allineamento

Le strutture della teoria dei giochi discusse in questa appendice — equilibri, mechanism design, regret minimization — sono i fondamenti formali per l'analisi del problema dell'allineamento che il prossimo saggio affronterà sistematicamente. Anticipiamo qui i tre nodi principali di quella trattazione.

Il primo è il problema del principal-agent: la relazione tra un sistema AI (l'agente) e i suoi progettisti o utilizzatori (il principale) può essere modellata come un gioco di informazione asimmetrica dove l'agente ha accesso a informazione privata — la propria distribuzione appresa di credenze e preferenze — che il principale non può osservare direttamente. Il mechanism design fornisce la struttura formale per progettare sistemi di addestramento e supervisione che inducano comportamenti allineati sotto questa asimmetria informativa.

Il secondo è il problema della cooperazione multi-agente: in scenari con più agenti AI che interagiscono — o con agenti AI che interagiscono con agenti umani — la teoria dei giochi predice che gli agenti razionali auto-interessati possono convergere verso equilibri inefficienti. La progettazione di strutture di incentivi che favoriscano la cooperazione senza richiedere altruismo — nel senso del mechanism design — è un problema aperto di grande importanza pratica.

Silvano Salvador

Il terzo è il problema dell'aggregazione delle preferenze: quando più agenti umani hanno preferenze diverse, la progettazione di un sistema AI che le aggreghi in modo equo incontra le impossibilità fondamentali del teorema di Arrow e del teorema di Gibbard-Satterthwaite. Non esiste nessun meccanismo di aggregazione delle preferenze che sia simultaneamente equo, efficiente e immune alla manipolazione strategica — una conclusione che ha implicazioni profonde per la progettazione di sistemi AI democratici e multi-stakeholder.

Appendice Tecnica II

Famiglie esponenziali, geometria duale e la struttura unificante delle distribuzioni parametriche

Nota editoriale. Questa appendice sviluppa la struttura formale delle famiglie esponenziali e la geometria duale che ne emerge, colmando un vuoto tra la trattazione del corpo principale e l'apparato della geometria dell'informazione sviluppato nell'Appendice VI del terzo saggio. I risultati qui derivati — in particolare la struttura dei potenziali coniugati, la connessione duale e-m e la proiezione di Pythagoras — sono le fondamenta su cui il Saggio III costruisce la teoria delle varietà statistiche applicata all'ottimizzazione delle policy. La lettura di questa appendice è consigliata prima di affrontare quella.

A.II.1 — La famiglia esponenziale: definizione canonica e struttura sufficiente

Una famiglia di distribuzioni $\{p(x;\,\theta) : \theta \in \Theta \subseteq \mathbb{R}^d\}$ è una famiglia esponenziale in forma canonica se la densità rispetto a una misura dominante μ ha la forma:

$$p(x;\,\theta) = h(x)\,\exp(\theta^{\mathsf{T}}\,T(x) - A(\theta))$$

Silvano Salvador

dove θ è il parametro naturale, $T : X \to \mathbb{R}^d$ è la statistica sufficiente, $h : X \to \mathbb{R}_+$ è la misura base, e $A : \Theta \to \mathbb{R}$ è la funzione cumulante o log-partizione:

$$A(\theta) = \log \int h(x) \exp(\theta^T T(x))\, d\mu(x)$$

definita su $\Theta = \{\theta : A(\theta) < \infty\}$, che è sempre convesso per la log-convessità dell'integranda. La funzione A determina completamente la struttura statistica della famiglia: derivando sotto il segno di integrale,

$$\partial A/\partial\theta_i = E_\theta[T_i(X)] =: \eta_i, \quad \partial^2 A/\partial\theta_i\partial\theta = \mathrm{Cov}_\theta[T_i(X), T(X)]$$

Il vettore $\eta = \nabla_\theta A(\theta)$ è il parametro dei momenti; la matrice hessiana $\nabla^2 A(\theta)$ è semidefinita positiva, confermando la convessità di A. Per famiglie non degeneri (componenti di T affinemente indipendenti), A è strettamente convessa e la matrice di Fisher coincide con questa hessiana:

$$G(\theta)_i = \partial^2 A/\partial\theta_i\partial\theta = \mathrm{Cov}_\theta[T_i(X), T(X)]$$

La geometria della famiglia esponenziale è dunque la geometria della funzione cumulante A: la curvatura di A è la curvatura dello spazio statistico.

Il teorema di Pitman-Koopman-Darmois completa il quadro affermando il viceversa: se una famiglia di distribuzioni ammette una statistica sufficiente di dimensione finita indipendentemente dalla dimensione campionaria, allora — sotto condizioni di regolarità — è una famiglia esponenziale. Le famiglie esponenziali sono pertanto la struttura parametrica minimale compatibile con la comprimibilità dell'informazione campionaria: esse sono le uniche famiglie in cui il campione di n osservazioni può essere sempre riassunto in un vettore di dimensione fissa.

Silvano Salvador

Esempi canonici nella forma naturale: la bernoulli con $\theta = \log(p/(1-p))$ e $A(\theta) = \log(1+e^{\theta})$; la gaussiana univariata con $\theta = (\mu/\sigma^2, -1/(2\sigma^2))$, $T(x) = (x, x^2)$ e $A(\theta) = -\theta_1^2/(4\theta_2) - \tfrac{1}{2}\log(-2\theta_2)$; la categorica con $\theta_i = \log(p_i/p)$ e $A(\theta) = \log(1 + \sum_i e^{\theta_i})$; la distribuzione di Dirichlet con $\theta = \alpha - 1$ e $A(\theta) = \sum \log \Gamma(\theta +1) - \log \Gamma(\sum (\theta +1))$.

A.II.2 — Il potenziale coniugato e la dualità di Legendre

La struttura di dualità emerge dalla trasformata di Legendre della funzione cumulante. Il potenziale coniugato è:

$$A^*(\eta) = \sup_{\theta \in \Theta} [\theta^T \eta - A(\theta)]$$

Per A strettamente convessa e differenziabile, il supremo è raggiunto nel θ^* tale che $\nabla A(\theta^*) = \eta$, da cui $A^*(\eta) = (\theta^*)^T \eta - A(\theta^*)$. Il potenziale coniugato ha un'interpretazione statistica precisa: sostituendo $\eta = E_\theta[T(X)]$ si verifica che

$$A^*(\eta) = -H(p(\cdot;\theta))$$

a meno di costanti dipendenti dalla misura base — A^* è l'entropia negativa della distribuzione corrispondente. Il segno è coerente con la convessità: A^* è convessa e l'entropia è concava.

Le relazioni di dualità seguono dall'involuzione della trasformata di Legendre:

$$\nabla A(\theta) = \eta, \quad \nabla A^*(\eta) = \theta, \quad A + A^* = \theta^T \eta$$

La coppia (A, A^*) definisce una struttura duale sullo spazio delle distribuzioni: ogni distribuzione ha una doppia rappresentazione, in termini del parametro naturale θ o del parametro dei momenti $\eta = \nabla A(\theta)$, collegati biunivocamente. Le due funzioni sono i potenziali della dualità, e la loro somma è il prodotto scalare delle due coordinate — una relazione che riecheggia la dualità di Legendre della meccanica classica

Silvano Salvador

(hamiltoniana e lagrangiana) e della termodinamica (energia libera di Helmholtz e di Gibbs).

A.II.3 — Le divergenze di Bregman e la KL come caso speciale

Per una funzione convessa differenziabile $\varphi : \Omega \to \mathbb{R}$, la divergenza di Bregman associata è:

$$D_\varphi(p \,\|\, q) = \varphi(p) - \varphi(q) - \nabla\varphi(q)^T(p - q) \geq 0$$

con uguaglianza se e soltanto se $p = q$. Essa misura il gap tra $\varphi(p)$ e la sua approssimazione lineare in q valutata in p — sempre non negativa per convessità.

La divergenza KL è la divergenza di Bregman generata dall'entropia negativa $\varphi(p) = \sum_x p(x) \log p(x)$:

$$D_{KL}(P \,\|\, Q) = D_\varphi(p \,\|\, q)$$

Per le famiglie esponenziali la KL ha una forma esplicita tramite il potenziale coniugato:

$$D_{KL}(p(\cdot;\theta_1) \,\|\, p(\cdot;\theta_2)) = A(\theta_2) - A(\theta_1) - \nabla A(\theta_1)^T(\theta_2 - \theta_1) = D_A(\theta_2 \,\|\, \theta_1)$$

Questa è la divergenza di Bregman generata da A, valutata nell'ordine (θ_2, θ_1). La simmetria crociata è il risultato fondamentale: la KL nella parametrizzazione naturale è la divergenza di Bregman di A nell'ordine invertito dei due argomenti. Nel parametro dei momenti vale l'analogo con A^*:

$$D_{KL}(p(\cdot;\theta_1) \,\|\, p(\cdot;\theta_2)) = D_{\{A^*\}}(\eta_1 \,\|\, \eta_2)$$

Silvano Salvador

La stessa quantità ha due rappresentazioni come divergenza di Bregman nei due sistemi di coordinate duali — una manifestazione concreta della dualità (A, A*).

A.II.4 — Il teorema di Pythagoras generalizzato e l'ottimizzazione variazionale

Il risultato geometrico più importante per l'inferenza variazionale è il teorema di Pythagoras generalizzato di Amari. Siano p, q, r tre distribuzioni in una famiglia esponenziale tali che la geodetica m da q a r sia e-ortogonale a quella da p a q in p. Allora:

$$D_{KL}(q \| r) = D_{KL}(q \| p) + D_{KL}(p \| r)$$

Questa è la versione informazionale del teorema di Pitagora, con la KL al posto della distanza euclidea quadratica. La condizione di e-ortogonalità è la generalizzazione non commutativa della perpendicolarità: p è la proiezione e-geodetica di q sulla sottovarietà m-planare contenente r.

Le implicazioni per l'inferenza variazionale sono immediate. Il minimizzatore di $D_{KL}(q \| p^*)$ su una famiglia approssimante Q:

$$q^* = \operatorname{argmin}_{q \in Q} D_{KL}(q \| p^*)$$

è la proiezione e-geodetica di p* sulla famiglia Q. Se Q è una famiglia esponenziale, questa proiezione equivale al matching dei momenti: q* è la distribuzione in Q con gli stessi momenti delle statistiche sufficienti di Q sotto p*.

Il teorema di Pythagoras garantisce che per qualsiasi $q \in Q$:

$$D_{KL}(q \| p^*) = D_{KL}(q \| q^*) + D_{KL}(q^* \| p^*)$$

Il gap tra q e p* si decompone in due contributi indipendenti: l'errore di ottimizzazione all'interno di Q (quanto q dista da q*) e l'errore di

Silvano Salvador

approssimazione irriducibile (quanto la migliore approssimazione in Q dista da p*). Questa decomposizione è il fondamento geometrico della misspecification variazionale: il secondo termine non può essere ridotto dall'ottimizzazione, ma soltanto dalla scelta di una famiglia Q più espressiva.

A.II.5 — L'algoritmo EM come alternanza di proiezioni duali

L'algoritmo Expectation-Maximization di Dempster, Laird e Rubin (1977) riceve una motivazione geometrica profonda nella geometria duale. Dato un modello con variabili latenti Z, si vuole massimizzare la log-verosimiglianza marginale log p(x; θ) mediante la funzione ausiliaria:

$Q(\theta, \theta_old) = E_\{Z|x,\theta_old\}[\log p(x, Z; \theta)]$

L'EM alterna:

E-step: calcola $q^*(Z) = p(Z \mid x; \theta_old)$ — la e-proiezione della distribuzione congiunta sulla sottovarietà compatibile con le osservazioni x

M-step: aggiorna $\theta_new = \text{argmax}_\theta\ Q(\theta, \theta_old)$ — la m-proiezione di $q^*(Z)\cdot p(x|Z;\theta)$ sulla famiglia parametrica

L'EM è dunque una sequenza di proiezioni alternate tra due sottovarietà ortogonali: la sottovarietà empirica (le distribuzioni compatibili con x) e la famiglia parametrica. La convergenza a un punto fisso è garantita dal teorema di Pythagoras: ogni proiezione non aumenta la divergenza KL dal punto fisso.

Il VAE è l'EM con e-proiezione approssimata: l'encoder $q_\varphi(z|x)$ approssima la proiezione e-geodetica esatta $p^*(z|x)$. Il gap dell'ELBO rispetto alla log-verosimiglianza vera è precisamente il termine di

Silvano Salvador

Pythagoras $D_{KL}(q_\varphi(z|x) \parallel p^*(z|x))$ — l'errore dell'approssimazione rispetto alla proiezione ottimale.

A.II.6 — Connessioni duali e geodetiche

Le due famiglie di geodetiche — e-geodetiche e m-geodetiche — descrivono i due modi naturali di muoversi nello spazio delle distribuzioni.

Le e-geodetiche tra $p(\cdot;\theta_0)$ e $p(\cdot;\theta_1)$ hanno la forma:

$$p_t(x) \propto p_0(x)^{1-t} \cdot p_1(x)^t = h(x) \exp(((1-t)\theta_0 + t\theta_1)^T T(x))$$

ovvero interpolazione lineare nel parametro naturale θ. Il parametro naturale del percorso è $(1-t)\theta_0 + t\theta_1$ — interpretazione geometrica della convessità di Θ.

Le m-geodetiche tra le stesse distribuzioni hanno forma:

$$p_t(x) = (1-t)\, p_0(x) + t\, p_1(x)$$

ovvero miscela lineare. L'interpolazione lineare nel parametro dei momenti η corrisponde alla miscela lineare delle densità. La divergenza KL è asimmetrica: il costo della e-geodetica da p_0 a p_1 è $D_{KL}(p_1 \parallel p_0)$, quello della m-geodetica è $D_{KL}(p_0 \parallel p_1)$. Le due divergenze misurano il costo di percorrere geodetiche in sistemi di connessioni duali opposti — il significato geometrico dell'asimmetria della KL discussa nel corpo principale.

A.II.7 — Famiglie esponenziali curve e curvatura statistica

Le famiglie esponenziali curve sono sottofamiglie di una famiglia piena ottenute parametrizzando il parametro naturale tramite una mappa differenziabile $\theta(\xi) : \Xi \to \Theta$ con $\dim(\Xi) < \dim(\Theta)$. La famiglia curva è $\{p(x; \theta(\xi)) : \xi \in \Xi\}$.

Silvano Salvador

La matrice di Fisher nel parametro ridotto è:

$$G_\xi(\xi)_i\square = (\partial\theta/\partial\xi_i)^T\, G_\theta(\theta(\xi))\, (\partial\theta/\partial\xi\,)$$

La curvatura della varietà $M = \{\theta(\xi) : \xi \in \Xi\}$ immersa in Θ — la curvatura statistica di Efron (1975) — determina le correzioni di secondo ordine all'inferenza: quanto le stime di massima verosimiglianza sulla famiglia curva si discostino da quelle sulla piena.

Le reti neurali sono famiglie esponenziali curve: la funzione $w \mapsto \theta(w)$ che mappa i pesi nelle distribuzioni determina la curvatura statistica del modello. Reti più profonde hanno genericamente curvatura statistica più elevata, il che spiega la sensibilità dell'addestramento alle condizioni iniziali: la varietà curva M può avere regioni di curvatura molto eterogenea, rendendo l'ottimizzazione nella metrica dei pesi inadeguata rispetto alla geometria intrinseca della varietà.

A.II.8 — Il softmax come famiglia esponenziale e la struttura dell'attenzione

La distribuzione softmax sui logit $\theta_i = q^T k_i/\sqrt{d}$ è la distribuzione categorica nella sua parametrizzazione esponenziale: la massima entropia soggetta al vincolo che il logaritmo delle probabilità sia lineare nei logit. La funzione cumulante è $A(\theta) = \log \sum_i \exp(\theta_i)$ — il log-sum-exp. La matrice di Fisher della categorica con probabilità p è:

$$G(p)_i\square = p_i(\delta_i\square - p\,)$$

che è la matrice di varianza-covarianza: positiva semidefinita, di rango n-1.

Il gradiente naturale dell'obiettivo rispetto ai logit θ produce aggiornamenti invarianti rispetto a riscalature delle logit che lasciano invariata la distribuzione di attenzione — una proprietà che il gradiente

Silvano Salvador

euclideo non soddisfa. La sensibilità alla scalatura dei logit nei transformer, e la necessità di scheduling del learning rate elaborato, riflettono l'eterogeneità della curvatura di Fisher lungo lo spazio dei logit: teste con attenzione concentrata (p vicina a un delta) hanno curvatura alta; teste con attenzione diffusa (p quasi uniforme) hanno curvatura bassa. L'Appendice VI del Saggio III sviluppa questa struttura sistematicamente per le distribuzioni delle azioni di un agente in un MDP.

A.II.9 — Normalità asintotica e connessione con Bernstein-von Mises

Per una famiglia esponenziale regolare con n osservazioni i.i.d. $x_1,...,x \sim p(\cdot;\theta_0)$, lo stimatore MLE soddisfa:

$\sqrt{n}\,(\hat{\theta} - \theta_0) \to N(0, G(\theta_0)^{-1})$ in distribuzione

Questa è la normalità asintotica nella metrica di Fisher: la varianza asintotica è l'inverso della matrice di informazione, il limite di Cramér-Rao raggiunto per la sufficienza di T(X). Nella geometria duale, la normalità asintotica del MLE significa che la distribuzione a posteriori (per prior regolari) converge a una gaussiana nella metrica di Fisher centrata su $\hat{\theta}$ — il teorema di Bernstein-von Mises per famiglie esponenziali. La convergenza è esponenzialmente rapida in n.

La connessione con il VAE è precisa: l'encoder approssima questa distribuzione a posteriori con una gaussiana diagonale. L'approssimazione è buona quando la vera posteriore è quasi gaussiana — il regime di molte osservazioni o prior informativi — e si deteriora nei regimi di poca informazione o distribuzione a posteriori multimodale, esattamente i casi in cui posterior collapse e distorsioni dell'ELBO sono più marcate. La geometria duale chiarisce perché: una distribuzione

Silvano Salvador

multimodale non è approssimabile con una gaussiana senza perdita sostanziale della struttura topologica — e questa perdita è irriducibile per il teorema di Pythagoras, indipendentemente dall'ottimizzazione.

Appendice Tecnica III

Ottimizzazione stocastica, adattività e la geometria della convergenza nei modelli profondi

Nota editoriale. Questa appendice sviluppa la matematica dell'ottimizzazione stocastica nei modelli profondi — la struttura formale che rende possibile l'addestramento dei modelli discussi nel corpo principale. Il materiale è organizzato in tre strati: la convergenza della discesa del gradiente stocastico e le sue garanzie asintotiche, la famiglia degli ottimizzatori adattativi (AdaGrad, Adam, RMSProp) come approssimazioni del gradiente naturale, e le tecniche di regolarizzazione implicita ed esplicita che producono la generalizzazione. Questo materiale è il ponte tra la teoria dell'ottimizzazione convessa classica — la cui struttura è garantita da teoremi di convergenza globale — e la realtà dell'addestramento profondo, dove la non-convessità è la norma e la comprensione teorica è incompleta.

A.III.1 — Discesa del gradiente e regimi di convergenza

Sia $f : \mathbb{R}^d \to \mathbb{R}$ una funzione di perdita differenziabile. L'aggiornamento della discesa del gradiente con passo fisso η è:

$$w_{t+1} = w - \eta \nabla f(w)$$

La convergenza dipende dalle proprietà di f. Per f L-liscia — con $\| \nabla f(w) - \nabla f(v) \| \leq L \|w - v\|$ per ogni w, v — la scelta $\eta \leq 1/L$ garantisce:

$$f(w_{t+1}) \leq f(w) - (\eta/2) \| \nabla f(w) \|^2$$

Silvano Salvador

Sommando su t e usando la non-negatività di f (o la limitatezza inferiore):

$$\min_{t\le T} \|\nabla f(w)\|^2 \le 2(f(w_0) - f^*) / (\eta T)$$

Questo bound garantisce convergenza a un punto stazionario (gradiente zero) al tasso $O(1/T)$, indipendentemente dalla non-convessità globale di f. La convergenza al minimo globale richiede convessità aggiuntiva.

Per f μ-fortemente convessa e L-liscia, il rapporto di condizionamento $\varkappa = L/\mu$ governa la convergenza esponenziale:

$$f(w) - f^* \le (1 - \mu\eta)^t (f(w_0) - f^*)$$

che con $\eta = 1/L$ diventa $(1 - \mu/L)^t (f(w_0) - f^*)$. Per problemi ben condizionati ($\varkappa$ piccolo) la convergenza è rapida; per problemi mal condizionati ($\varkappa$ grande, tipico nelle reti profonde) il numero di iterazioni per ridurre l'errore di un fattore ε è $O(\varkappa \log(1/\varepsilon))$ — lineare nel condizionamento.

Il gradiente stocastico (SGD) sostituisce $\nabla f(w)$ con un estimatore $\hat{g}$ costruito su un mini-batch $B \subset \{1,...,n\}$:

$$\hat{g} = (1/|B|) \sum_{i \in B} \nabla f_i(w), \quad E[\hat{g} \mid w] = \nabla f(w)$$

La varianza del gradiente stocastico, $\sigma^2 = E[\|\hat{g} - \nabla f(w)\|^2]$, introduce un termine di rumore irriducibile che impone il decremento del passo η per la convergenza. Con passo decrescente $\eta = \eta_0/\sqrt{t}$, l'SGD converge a un punto stazionario al tasso $O(\sigma/\sqrt{T})$ per funzioni non convesse L-lisce:

$$E[\min_{t\le T} \|\nabla f(w)\|^2] \le 2(f(w_0) - f^*) / (\eta_0\sqrt{T}) + \eta_0\sigma^2 L$$

Il termine $\sigma^2 L$ non si annulla mai: esso è il costo inevitabile della stocasticità, la varianza introdotta dal mini-batching che non può essere eliminata con più iterazioni, soltanto ridotta aumentando $|B|$ (e dunque il costo computazionale per iterazione).

Silvano Salvador

A.III.2 — AdaGrad e l'adattività coordinata

AdaGrad di Duchi, Hazan e Singer (2011) modifica SGD accumulando la somma dei quadrati dei gradienti per coordinate:

$G_t = \sum_{s=1}^{t} \hat{g}_s \hat{g}_s^\top$ (matrice diagonale: $G_{t,ii} = \sum_s \hat{g}_{s,i}^2$)

$w_{t+1} = w_t - \eta \, (\mathrm{diag}(G_t) + \varepsilon I)^{-1/2} \hat{g}_t$

Il denominatore $(G_{t,ii})^{1/2}$ scala adattativamente il passo per la coordinata i in modo inversamente proporzionale alla radice della somma storica dei quadrati dei gradienti su quella coordinata. Le coordinate con gradienti storicamente grandi ricevono passi piccoli; quelle con gradienti piccoli ricevono passi grandi.

Il bound sul regret di AdaGrad è:

$$\sum_{t=1}^{T} [f(w_t) - f(w^*)] \leq \eta/(2\varepsilon) \, \|w_0 - w^*\|^2 + \eta \sum_{i=1}^{d} (\sum_{t=1}^{T} \hat{g}_{t,i}^2)^{1/2}$$

Il secondo termine è controllato dalla geometria del problema: se poche coordinate hanno gradienti grandi e molte hanno gradienti piccoli (sparsità nel gradiente), AdaGrad converge molto più velocemente di SGD con passo fisso. Questo è il regime tipico dell'apprendimento da embedding sparsi — word vectors, item embeddings — dove soltanto una piccola frazione dei parametri riceve aggiornamenti non nulli per ogni campione.

Il limite di AdaGrad è la monotonica crescita di $G_{t,ii}$: il passo adattativo per la coordinata i decresce come $t^{-1/2}$ anche se i gradienti recenti sono piccoli. In presenza di struttura non stazionaria — come nell'addestramento dei modelli profondi, dove la loss landscape cambia significativamente durante l'addestramento — la memoria infinita di AdaGrad diventa un limite.

Silvano Salvador

A.III.3 — RMSProp, Adam e la memoria esponenziale

RMSProp di Hinton (2012) e Adam di Kingma e Ba (2015) risolvono il problema della memoria infinita di AdaGrad introducendo una media mobile esponenziale dei quadrati dei gradienti.

RMSProp mantiene la stima:

$v = \beta v_{-1} + (1-\beta)\, \hat{g}^2$ (elemento per elemento)

e aggiorna:

$w_{t+1} = w - \eta\, \hat{g} / (\sqrt{v} + \varepsilon)$

La media esponenziale con fattore di decay $\beta \in (0,1)$ "dimentica" i gradienti passati con ritmo geometrico, consentendo all'adattività di rispondere ai cambiamenti nel paesaggio di perdita.

Adam combina RMSProp con un momento del primo ordine. Mantiene le stime:

$m = \beta_1 m_{-1} + (1-\beta_1)\, \hat{g}$ (primo momento: media mobile del gradiente)

$v = \beta_2 v_{-1} + (1-\beta_2)\, \hat{g}^2$ (secondo momento: media mobile del gradiente quadratico)

con correzione del bias di inizializzazione:

$\hat{m} = m / (1 - \beta_1^t), \quad \hat{v} = v / (1 - \beta_2^t)$

e aggiornamento:

$w_{t+1} = w - \eta\, \hat{m} / (\sqrt{\hat{v}} + \varepsilon)$

La correzione del bias è necessaria perché $m_0 = v_0 = 0$: senza correzione, le prime iterazioni sovrastimerebbero sistematicamente la norma del gradiente — le stime iniziali sarebbero polarizzate verso zero per la memoria esponenziale su una finestra ancora breve.

Silvano Salvador

Adam approssima — in modo coordinata per coordinata, non globale — la discesa del gradiente naturale: il denominatore $\sqrt{\hat{v}_t}$ stima la deviazione standard del gradiente, una misura locale della curvatura della funzione di perdita. L'approssimazione è rozza rispetto alla matrice di Fisher completa (che è $d \times d$), ma computazionalmente trattabile ($O(d)$ per iterazione contro $O(d^2)$ del gradiente naturale pieno) e sorprendentemente efficace in pratica.

Le garanzie teoriche di convergenza di Adam in contesti non convessi rimangono incomplete: Reddi, Kale e Kumar (2018) hanno dimostrato che Adam può non convergere anche su problemi convessi con sequenze di gradienti sfavorevoli, a causa della non-monotonia del passo adattativo con $\beta_2 < 1$. AMSGrad di Reddi et al. corregge questo problema sostituendo $\hat{v}_t$ con $\max(\hat{v}_t, \hat{v}_{t-1})$, garantendo la monotonia del passo e ripristinando le garanzie di convergenza al costo di prestazioni empiriche leggermente inferiori.

A.III.4 — Regolarizzazione esplicita: weight decay, dropout e normalizzazione

Le tecniche di regolarizzazione modificano l'obiettivo di addestramento o l'architettura del modello per ridurre l'overfitting. Le tre più importanti hanno strutture matematiche diverse.

Il weight decay aggiunge un termine di penalizzazione L2 all'obiettivo:

$$L_{reg}(w) = L(w) + (\lambda/2)\,\|w\|^2$$

Il gradiente è $\nabla L_{reg} = \nabla L + \lambda w$, dunque l'aggiornamento SGD con weight decay è:

$$w_{t+1} = (1 - \eta\lambda)\, w_t - \eta\, \nabla L(w_t)$$

Silvano Salvador

Il fattore $(1 - \eta\lambda) < 1$ decrementa i pesi a ogni passo verso zero, penalizzando le soluzioni con norma elevata. Nella prospettiva bayesiana, il weight decay corrisponde a un prior gaussiano isotropo sui pesi: $p(w) \propto \exp(-(\lambda/2)\|w\|^2)$, e l'ottimizzazione dell'obiettivo regolarizzato è la massimizzazione a posteriori (MAP) con questo prior. La corrispondenza è esatta: $\lambda = \sigma_prior^{-2}$ per un prior $N(0, \sigma_prior^2 I)$.

Il dropout di Srivastava e colleghi (2014) disattiva casualmente ogni neurone con probabilità p durante l'addestramento, equivalentemente campiona una sottorete diversa per ogni mini-batch. L'obiettivo effettivo è:

$$L_dropout(w) = E_\xi[L(w \odot \xi)]$$

dove ξ è una maschera binaria con $E[\xi_i] = 1-p$. L'effetto regolarizzante del dropout è molteplice: promuove rappresentazioni distribuite ridondanti (nessun neurone può affidarsi esclusivamente ad altri neuroni per completare un compito), introduce rumore nell'ottimizzazione (SGD stocastico sulla sottorete campionata), e può essere interpretato come un meccanismo di model averaging implicito (ogni maschera definisce un modello, e il modello finale è l'ensemble su tutte le maschere).

La normalizzazione a batch (Ioffe e Szegedy, 2015) normalizza le attivazioni di ogni layer per ogni mini-batch:

$$\mu_B = (1/|B|) \sum_\{i \in B\} x_i, \quad \sigma_B^2 = (1/|B|) \sum_\{i \in B\} (x_i - \mu_B)^2$$

$$\hat{x_i} = (x_i - \mu_B) / \sqrt{(\sigma_B^2 + \varepsilon)}, \quad y_i = \gamma \hat{x_i} + \beta$$

con parametri apprendibili γ e β. La normalizzazione riduce il covariate shift interno — la variazione della distribuzione delle attivazioni durante l'addestramento — permettendo passi di apprendimento più grandi senza instabilità. Il suo effetto regolarizzante è in parte dovuto al rumore introdotto dalla stima stocastica di μ_B e σ_B sul mini-batch.

Silvano Salvador

A.III.5 — Regolarizzazione implicita e l'implicit bias di SGD

I modelli sovraparametrizzati hanno tipicamente infiniti punti di perdita nulla sull'insieme di addestramento: la loss landscape è degenere, con un continuum di soluzioni globali. L'ottimizzatore seleziona tra queste soluzioni in base alla propria traiettoria — un effetto di regolarizzazione implicita non esplicitamente codificato nell'obiettivo.

Per i modelli lineari con SGD e passo infinitesimo (gradient flow), il risultato è noto con precisione. Il gradiente flow sul problema di regressione con perdita quadratica, partendo da $w_0 = 0$, converge alla soluzione di minima norma euclidea:

$$w^* = \text{argmin}_{w : Xw = y} \|w\|^2$$

indipendentemente dall'inizializzazione in zero. Questa è la soluzione della pseudoinversa $w^* = X^T(XX^T)^{-1}y$. Per la classificazione con perdita esponenziale o logistica, il gradient flow converge invece al classificatore di massimo margine — il separatore degli SVM:

$$w^* = \text{argmax}_{w : y_i w^T x_i \geq 1} \|w\|^{-1}$$

ovvero la direzione in cui il margine è massimizzato. Il bias verso la minima norma (regressione) e il massimo margine (classificazione) sono conseguenze della geometria euclidea dell'ottimizzazione e dell'inizializzazione a zero, non di scelte esplicite del programmatore.

Per le reti neurali profonde l'implicit bias è più complesso e meno compreso. Gunasekar e colleghi (2018) hanno dimostrato che il gradient flow su reti lineari profonde (prodotti di matrici) converge verso soluzioni che minimizzano la norma nucleare del prodotto — equivalente a favorire soluzioni di basso rango. Questo ha interpretazioni computazionali: le reti profonde hanno un bias intrinseco verso le rappresentazioni di bassa complessità statistica, e questo bias dipende

Silvano Salvador

dalla profondità dell'architettura. Il collegamento con la teoria MDL della Parte Quarta è strutturale: l'addestramento con SGD non è neutro rispetto alla complessità del modello, ma favorisce implicitamente le soluzioni più semplici compatibili con i dati di addestramento — esattamente il principio di Occam nella formulazione MDL.

A.III.6 — Loss landscape, punti critici e la struttura degli ottimi locali

La funzione di perdita delle reti neurali profonde è genericamente non convessa, con un numero esponenziale di minimi locali, punti di sella e regioni piatte. Comprendere la struttura di questa landscape è essenziale per spiegare perché SGD raggiunga soluzioni di buona qualità nonostante la non-convessità globale.

Un primo risultato è che per le reti suficientemente sovraparametrizzate, i minimi locali tendono ad avere valori di perdita comparabili ai minimi globali. Du e colleghi (2019) e Allen-Zhu e colleghi (2019) hanno dimostrato che per reti a due strati con ampiezza sufficiente (proporzionale a n^6 per n punti di addestramento, un bound largamente sub-ottimale rispetto alla pratica), il gradiente flow converge a un minimo globale dalla maggior parte delle inizializzazioni casuali. Il regime di larghezza richiesto dai bound teorici è molto più ampio di quello usato in pratica: le reti pratiche raggiungono minimi globali con architetture molto più strette, suggerendo che la struttura reale della landscape sia più benigna di quanto catturato dai bound attuali.

I punti di sella — punti stazionari con direzioni di curvatura sia positiva che negativa — sono più problematici dei minimi locali per SGD. Tuttavia, il rumore del gradiente stocastico aiuta SGD a sfuggire dai punti di sella: il gradiente di un campione casuale ha con alta probabilità una componente non nulla nella direzione di fuga dal punto di sella. Il

Silvano Salvador

risultato di Jin e colleghi (2017) formalizza questo: SGD con rumore appropriato sfugge ai punti di sella approssimativi in tempo polinomiale nel numero di parametri, convergendo eventualmente a un minimo locale che è anche un minimo locale stretto (con tutte le curvature positive).

La struttura delle regioni piatte — regioni dello spazio dei parametri dove la perdita varia poco — ha implicazioni per la generalizzazione. Hochreiter e Schmidhuber (1997) hanno proposto che i minimi "piatti" (con bassa curvatura, misurata dall'autovalore massimo dell'hessiana) generalizzino meglio dei minimi "stretti" (alta curvatura). Questa intuizione è correlata alla relazione tra curvatura e PAC-Bayes bounds: la misura di Boltzmann sulla landscape di perdita pesa i parametri inversamente alla loro curvatura, e le regioni piatte sono quelle che ricevono più massa sotto questa misura. Keskar e colleghi (2017) hanno mostrato empiricamente che i batch grandi tendono a convergere a minimi stretti con peggiore generalizzazione rispetto ai batch piccoli, confermando la relazione tra rumore dell'ottimizzatore e piattezza del minimo raggiunto.

A.III.7 — Il learning rate scheduling: warmup, cosine e one-cycle

Il passo di apprendimento η è l'iperparametro più critico nell'addestramento dei modelli profondi. Passi troppo grandi causano instabilità o divergenza; passi troppo piccoli rallentano la convergenza. Le strategie di scheduling adattano η durante l'addestramento per bilanciare queste esigenze.

Il warmup lineare segue:

$\eta = \eta_max \cdot \min(t / T_warmup, 1)$

Silvano Salvador

Nelle fasi iniziali dell'addestramento, le stime del secondo momento v̂□ in Adam sono inaccurate (dominate dal bias di inizializzazione) e il denominatore può causare passi effettivi molto grandi o irregolari. Il warmup riduce il passo iniziale permettendo alle statistiche adattative di accumularsi prima che vengano sfruttate appieno. Ma la vera giustificazione è più sottile: nelle primissime iterazioni, i parametri sono lontani da qualsiasi struttura significativa della loss landscape, e la curvatura locale non è rappresentativa della curvatura nelle regioni di interesse. Muoversi lentamente nelle prime fasi riduce il rischio di convergere prematuramente verso regioni sub-ottimali basandosi su curvature locali fuorvianti.

Il cosine annealing decrementa η da η_max a η_min seguendo un mezzo coseno:

$$\eta = \eta_min + \tfrac{1}{2}(\eta_max - \eta_min)(1 + \cos(\pi t/T))$$

Questo schedulatore è particolarmente efficace perché il decremento è lento nelle fasi intermedie — permettendo all'ottimizzatore di esplorare ampie regioni della landscape — e rapido verso la fine — forzando la convergenza verso un minimo locale stretto prima della fine dell'addestramento. La combinazione cosine con warmup (warmup cosine schedule) è diventata lo standard de facto per i grandi modelli linguistici.

Il one-cycle policy di Smith e Touvron combina warmup e decremento in un unico ciclo: il passo cresce linearmente fino a η_max, poi decresce seguendo un annealing coseno fino a η_min con $\eta_min << \eta_max/1000$. Il picco del passo nel mezzo dell'addestramento ha una giustificazione geometrica: nella fase intermedia il modello si trova tipicamente vicino a regioni di buona struttura nella landscape, e un

Silvano Salvador

passo grande permette di esplorare efficacemente queste regioni senza restare intrappolati in minimi locali sub-ottimali.

A.III.8 — Batch size, generalizzazione e la connessione con la teoria PAC-Bayes

La scelta della batch size ha conseguenze sia computazionali sia statistiche. Computazionalmente, batch grandi sfruttano meglio il parallelismo delle GPU e riducono il numero di aggiornamenti per epoca; statisticamente, batch grandi riducono la varianza del gradiente stimato ma aumentano il rischio di convergenza verso minimi stretti.

La connessione con la generalizzazione è formalizzata dalla teoria PAC-Bayes. Un bound PAC-Bayes per una distribuzione posteriore Q sui parametri (con prior P) afferma che, con alta probabilità su il campione di addestramento, la perdita di generalizzazione soddisfa:

$$E_{w\sim Q}[L_{test}(w)] \leq E_{w\sim Q}[L_{train}(w)] + \sqrt{\left(D_{KL}(Q \,\|\, P) + \log(1/\delta)\right) / (2n)}$$

Il termine di complessità è $D_{KL}(Q \,\|\, P)$: quanto la distribuzione a posteriori appresa si discosta dal prior. Per la scelta $Q = N(w^*, \lambda I)$ — una gaussiana centrata sui pesi appresi con varianza inversamente proporzionale alla curvatura locale — il bound si riduce al rapporto tra la piattezza del minimo (curvatura bassa ↔ λ grande ↔ KL piccola) e la dimensione campionaria n. I minimi piatti generalizzano meglio perché corrispondono a posteriori P più concentrate sul prior, dunque a soluzioni con minore complessità nel senso PAC-Bayes — una formalizzazione rigorosa del principio di parsimonia di Occam e del criterio MDL della Parte Quarta.

Batch piccoli, producendo più rumore nell'ottimizzazione, tendono a convergere verso minimi più piatti: il rumore agisce come un

Silvano Salvador

meccanismo di selezione delle soluzioni a bassa complessità PAC-Bayes. Batch grandi riducono questo meccanismo selettivo, aumentando il rischio di convergere verso minimi stretti con peggiore generalizzazione. Questa connessione tra batch size, rumore dell'ottimizzatore, piattezza del minimo e generalizzazione è uno dei risultati teorici più solidi dell'ottimizzazione profonda contemporanea.

A.III.9 — Gradient clipping, instabilità e i modelli su lunga sequenza

Il gradient clipping è una tecnica fondamentale per l'addestramento di reti ricorrenti e transformer su sequenze lunghe. L'aggiornamento è:

$$w_{t+1} = w - \eta \hat{g} / \max(1, \|\hat{g}\|/c)$$

dove c è la soglia di clipping. Se $\|\hat{g}\| > c$, il gradiente viene riscalato a norma c; altrimenti è lasciato invariato.

Il gradient clipping risponde al problema dell'exploding gradient nelle reti profonde: quando la backpropagation attraversa molti strati o molti passi temporali, i gradienti possono crescere esponenzialmente per effetto della moltiplicazione di jacobiane. Per una rete ricorrente con pesi W e stato $h = \sigma(Wh_{-1} + Wx)$, il gradiente della perdita al passo T rispetto a h_0 è proporzionale a $(\partial h_T/\partial h_0) = \prod (\partial h / \partial h_{-1})$ — un prodotto di T matrici jacobiane. Se il valore spettrale maggiore di queste matrici è > 1, il prodotto cresce esponenzialmente in T.

Il gradient clipping non risolve il problema concettualmente — l'informazione sul gradiente viene troncata — ma lo rende praticamente gestibile: i passi di aggiornamento rimangono bounded, e l'addestramento è stabile. La versione per norma globale, descritta sopra, è teoricamente più motivata del clipping per coordinata perché preserva

Silvano Salvador

la direzione del gradiente: si riduce la lunghezza del passo ma non si modifica la direzione di discesa stimata.

Per i transformer, il gradient clipping è meno critico delle reti ricorrenti — la struttura dell'attenzione evita il propagation lungo catene temporali lunghe — ma rimane importante nelle prime fasi dell'addestramento quando le statistiche adattative di Adam non si sono stabilizzate. La combinazione warmup + gradient clipping nelle prime iterazioni è la prassi standard per i grandi modelli linguistici: warmup riduce l'ampiezza degli aggiornamenti, clipping ne limita la norma, e insieme assicurano che le primissime iterazioni non distruggano l'inizializzazione casuale con aggiornamenti catastrofici.

Appendice Tecnica IV

Diffusione stocastica e score matching: dimostrazioni complete, equazioni di Fokker-Planck e convergenza dei processi generativi

A.IV.1 — Processi stocastici e moti browniani: fondamenta

Un processo stocastico a tempo continuo $\{X_t\}_{t \geq 0}$ su $\mathbb{R}^n$ è una famiglia di variabili casuali indicizzate dal tempo, definite su uno spazio di probabilità $(\Omega, \mathcal{F}, \mathbb{P})$. La teoria dei processi di diffusione si fonda sul moto browniano standard, la cui costruzione rigorosa — dovuta a Wiener nel 1923 — richiede la specificazione di tre condizioni: $W_0 = 0$ quasi certamente; per ogni $0 \leq s < t$, l'incremento $W_t - W_s$ è distribuito come $\mathcal{N}(0, (t-s)\, I_n)$ e indipendente da $\mathcal{F}_s$, la σ-algebra generata da $\{W_r\}_{r \leq s}$; le traiettorie $t \mapsto W_t(\omega)$ sono continue quasi certamente.

Silvano Salvador

La varianza lineare nel tempo — $\mathbb{E}[\|W_t - W_s\|^2] = n(t-s)$ — codifica la natura diffusiva del processo: la distanza tipica percorsa in un intervallo di tempo τ scala come $\sqrt{\tau}$, non come τ. Questa proprietà distingue il moto browniano dai processi deterministici e riflette l'assenza di correlazioni tra incrementi non sovrapposti.

Un'equazione differenziale stocastica (SDE) di Itô della forma

$$dX_t = f(X_t, t)\,dt + G(X_t, t)\,dW_t$$

descrive un processo in cui il termine deterministico $f(X_t, t)$ — il drift — e il termine stocastico $G(X_t, t)\,dW_t$ — la diffusione — governano congiuntamente l'evoluzione. L'integrale stocastico $\int_0^T G(X_t, t)\,dW_t$ è definito nel senso di Itô come limite in probabilità di somme di Riemann con punti di valutazione all'estremo sinistro di ogni intervallo. Questa scelta — distinta dalla convenzione di Stratonovich, che usa il punto medio — ha implicazioni sulla formula di cambiamento di variabile.

La formula di Itô afferma che, per una funzione $\varphi : \mathbb{R}^n \times [0,T] \to \mathbb{R}$ sufficientemente regolare, il processo $\varphi(X_t, t)$ soddisfa:

$$d\varphi = \left(\partial_t \varphi + (\nabla_x \varphi)^T f + \tfrac{1}{2}\,\mathrm{tr}(GG^T \nabla_x^2 \varphi)\right) dt + (\nabla_x \varphi)^T G\,dW_t$$

Il termine aggiuntivo $\tfrac{1}{2}\,\mathrm{tr}(GG^T \nabla_x^2 \varphi)$ rispetto alla formula ordinaria della catena è la correzione stocastica di Itô, e riflette il fatto che le traiettorie browniane hanno variazione quadratica non nulla: $\sum (\Delta W_i)^2 \to t$ in probabilità quando il passo di discretizzazione tende a zero.

A.IV.2 — L'equazione di Fokker-Planck: derivazione

L'equazione di Fokker-Planck descrive l'evoluzione della densità di probabilità $p_t(x)$ del processo X_t governato dalla SDE di Itô. Essa è il duale dell'equazione di Kolmogorov all'indietro ed è fondamentale per l'analisi teorica dei modelli di diffusione.

Silvano Salvador

Sia $\varphi : \mathbb{R}^n \to \mathbb{R}$ una funzione test liscia a supporto compatto. La derivata temporale dell'aspettativa $\mathbb{E}[\varphi(X_t)]$ può essere calcolata in due modi: per la formula di Itô applicata a $\varphi(X_t)$, e per integrazione di $p_t(x)$ contro $\varphi(x)$.

Per la formula di Itô con $G = g(t)\, I$ (diffusione scalare isotropa, che è il caso rilevante per i modelli di diffusione):

$$d\mathbb{E}[\varphi(X_t)] = \mathbb{E}[\partial_t \varphi(X_t) + (\nabla_x \varphi(X_t))^T f(X_t,t) + \tfrac{1}{2} g(t)^2 \Delta_x \varphi(X_t)]\, dt$$

Per integrazione di p_t contro φ:

$$d/dt \int \varphi(x)\, p_t(x)\, dx = \int \varphi(x)\, \partial_t p_t(x)\, dx$$

Uguagliando le due espressioni e integrando per parti (assumendo che p_t e φ decadano all'infinito):

$$\int \varphi(x)\, \partial_t p_t(x)\, dx = \int \varphi(x)\, [-\mathrm{div}(f(x,t)\, p_t(x)) + \tfrac{1}{2} g(t)^2 \Delta p_t(x)]\, dx$$

Poiché φ è arbitraria, si ottiene l'equazione di Fokker-Planck:

$$\partial_t p_t(x) = -\mathrm{div}(f(x,t)\, p_t(x)) + \tfrac{1}{2} g(t)^2 \Delta p_t(x)$$

ovvero, in forma compatta con $D(t) = \tfrac{1}{2} g(t)^2$:

$$\partial_t p_t = -\mathrm{div}(f\, p_t) + D(t)\, \Delta p_t$$

Il primo termine — la divergenza del flusso deterministico — trasporta la densità lungo il campo di drift. Il secondo termine — il laplaciano ponderato per la diffusività $D(t)$ — spande la densità nello spazio, smussando le concentrazioni locali. La competizione tra questi due termini determina il comportamento a lungo termine del processo.

Per il caso del processo di diffusione forward utilizzato nei DDPM, con $f(x,t) = -\beta(t)/2 \cdot x$ e $g(t) = \sqrt{\beta(t)}$, l'equazione di Fokker-Planck diventa:

Silvano Salvador

$\partial_t p_t = \mathrm{div}(\beta(t)/2 \cdot x \cdot p_t) + \tfrac{1}{2}\beta(t)\,\Delta p_t$

$= \beta(t)/2 \cdot (\mathrm{div}(x\, p_t) + \Delta p_t)$

$= \beta(t)/2 \cdot (n\, p_t + x \cdot \nabla p_t + \Delta p_t)$

Questo è il generatore dell'equazione di Ornstein-Uhlenbeck, il cui punto fisso — la soluzione stazionaria $\partial_t p^* = 0$ — è la gaussiana standard $\mathcal{N}(0, I)$, come si verifica sostituendo $p^*(x) = (2\pi)^{-n/2} \exp(-\|x\|^2/2)$ nell'equazione.

A.IV.3 — La SDE reverse di Anderson: derivazione completa

Il risultato fondamentale di Anderson (1982) afferma che a ogni processo di diffusione forward corrisponde un processo di diffusione reverse che ha le stesse distribuzioni marginali ma scorre a ritroso nel tempo. Questo risultato è la base matematica dell'inversione del processo di diffusione nei modelli generativi.

Sia $\{X_t\}_{t \in [0,T]}$ il processo forward governato da:

$dX_t = f(X_t, t)\,dt + g(t)\,dW_t$

con distribuzione marginale $p_t(x)$ al tempo t. Definiamo il processo reverse $\{\bar{X}_t\}_{t \in [0,T]}$ con $\bar{X}_t := X_{T-t}$. Anderson dimostrò che questo processo soddisfa la SDE:

$d\bar{X}_t = [-f(\bar{X}_t, T-t) + g(T-t)^2\, \nabla_x \log p_{T-t}(\bar{X}_t)]\,dt + g(T-t)\,d\bar{W}_t$

dove $\bar{W}_t$ è un moto browniano standard rispetto alla filtrazione naturale di $\bar{X}_t$.

La derivazione si basa sulla formula di Girsanov e sull'analisi delle misure di probabilità sui percorsi. Qui ne offriamo una versione semplificata che cattura la struttura essenziale.

Silvano Salvador

Consideriamo la densità di transizione del processo forward, $p(x_t \mid x_s)$ per $s < t$. Per la SDE di Itô, questa densità soddisfa la Fokker-Planck vista come funzione di (x_t, t) con condizione iniziale δ_{x_s}. Per un processo con coefficiente di diffusione scalare, la relazione di reversibilità temporale tra le densità di transizione forward e reverse è data da:

$$p(x_s \mid x_t)\, p_t(x_t) = p(x_t \mid x_s)\, p_s(x_s)$$

Questa è semplicemente la regola di Bayes applicata alle densità di transizione. Il drift del processo reverse si ottiene da questa relazione scrivendo la SDE che genera la densità di transizione $p(x_s \mid x_t)$ come funzione del processo corrente.

Per la versione in tempo continuo, il drift reverse al tempo t (in notazione forward) è:

$$f_{\text{reverse}}(x, t) = -f(x, t) + g(t)^2\, \nabla_x \log p_t(x)$$

Il termine aggiuntivo $g(t)^2\, \nabla_x \log p_t(x)$ — la score function della distribuzione marginale al tempo t, scalata per $g(t)^2$ — ha un'interpretazione precisa: esso corregge il drift forward rimuovendo la componente che spinge la distribuzione verso la gaussiana standard, e aggiunge una componente che spinge verso le regioni di alta densità di p_t. Quando p_t è la gaussiana standard e $f(x,t) = -\beta(t)/2 \cdot x$, il termine correttivo è $-g(t)^2\, x$ e il drift reverse è zero: partendo dalla gaussiana, non c'è nessuna direzione privilegiata verso cui muoversi, coerentemente con il fatto che la gaussiana standard è il punto fisso del processo forward.

A.IV.4 — Score matching: la disuguaglianza di Fisher e la convergenza

Il legame tra score matching e la divergenza di Fisher — una grandezza distinta sia dalla divergenza KL sia dall'informazione di Fisher —

Silvano Salvador

chiarisce perché minimizzare la distanza tra score functions sia un obiettivo statisticamente fondato.

La divergenza di Fisher tra due distribuzioni P e Q con densità p e q è definita come:

$$J(P \| Q) = \mathbb{E}_P[\|\nabla_x \log p(x) - \nabla_x \log q(x)\|^2]$$

Essa misura la distanza media al quadrato tra le score functions delle due distribuzioni, dove la media è calcolata rispetto a P. A differenza della divergenza KL, la divergenza di Fisher è una quantità di secondo ordine nella distanza tra le distribuzioni: per distribuzioni vicine, essa è proporzionale alla distanza quadratica tra i parametri pesata dalla metrica di Fisher — la stessa connessione che collega la divergenza KL alla metrica di Fisher al secondo ordine.

L'obiettivo di score matching di Hyvärinen minimizza J(P_data ‖ P_model), ovvero la divergenza di Fisher dalla distribuzione dei dati alla distribuzione del modello. Hyvärinen dimostrò che, a meno di costanti indipendenti dal modello, questo equivale a minimizzare:

$$J_{implicit}(\theta) = \mathbb{E}_{p_{data}(x)}\bigl[\mathrm{tr}(\nabla_x s_\theta(x)) + \tfrac{1}{2}\|s_\theta(x)\|^2\bigr]$$

dove $s_\theta(x) = \nabla_x \log p(x; \theta)$ è la score function del modello. La derivazione per integrazione per parti procede componente per componente.

Per la i-esima componente del vettore gradiente:

$$\mathbb{E}_{p_{data}}[(s_\theta)_i(x) \cdot (s_{data})_i(x)]$$

$$= \int p_{data}(x)\, \partial_{x_i} \log p_\theta(x) \cdot \partial_{x_i} \log p_{data}(x)\, dx$$

$$= \int \partial_{x_i} \log p_\theta(x) \cdot \partial_{x_i} p_{data}(x)\, dx$$

Silvano Salvador

Integrando per parti rispetto a x_i (con termine al bordo nullo per ipotesi di decadimento):

$= -\int \partial^2_{x_i\, x_i} \log p_\theta(x) \cdot p_{data}(x)\, dx$

$= -\mathbb{E}_{p_{data}}[\partial^2_{x_i} \log p_\theta(x)]$

Sommando su tutte le componenti:

$\mathbb{E}_{p_{data}}[s_\theta(x)^T s_{data}(x)] = -\mathbb{E}_{p_{data}}[\mathrm{tr}(\nabla_x s_\theta(x))]$

Espandendo il quadrato nell'obiettivo originale e sostituendo:

$J(P_{data} \parallel P_{model}) = \mathbb{E}_{p_{data}}[\|s_\theta(x)\|^2 + 2\, s_\theta(x)^T s_{data}(x)] + \text{costante}$

$= \mathbb{E}_{p_{data}}[\|s_\theta(x)\|^2 - 2\, \mathrm{tr}(\nabla_x s_\theta(x))] + \text{costante}$

$= \mathbb{E}_{p_{data}}[\mathrm{tr}(\nabla_x s_\theta(x)) + \tfrac{1}{2} \|s_\theta(x)\|^2] \cdot 2 + \text{costante}$

confermando la forma implicita dell'obiettivo.

La difficoltà computazionale del termine $\mathrm{tr}(\nabla_x s_\theta(x))$ — la traccia dello Jacobiano della score function — per dimensioni elevate può essere affrontata con la stima di Hutchinson:

$\mathrm{tr}(A) = \mathbb{E}_{v \sim \mathcal{N}(0,I)}[v^T A v]$

da cui:

$\mathrm{tr}(\nabla_x s_\theta(x)) \approx v^T (\nabla_x s_\theta(x))\, v = v^T \nabla_x (s_\theta(x)^T v)$

per un campione casuale v. Questa stima ha varianza $O(n^2)$ per la traccia di matrici generali, ma può avere varianza molto inferiore quando la Jacobiana di s_θ è approssimativamente di basso rango — circostanza che si verifica in dati strutturati ad alta dimensionalità dove la score function varia significativamente soltanto lungo poche direzioni principali.

Silvano Salvador

A.IV.5 — Denoising score matching: equivalenza con score matching e stima della varianza

La dimostrazione formale di Vincent (2011) che il denoising score matching è equivalente allo score matching implicito sulla distribuzione smoothed richiede l'analisi della relazione tra la score function del processo di diffusione a diversi livelli di rumore.

Sia $q_\sigma(\tilde{x} \mid x) = \mathcal{N}(\tilde{x}; x, \sigma^2 I)$ il kernel di smoothing. La distribuzione smoothed è:

$$p_\sigma(\tilde{x}) = \int p_{data}(x)\, q_\sigma(\tilde{x} \mid x)\, dx$$

La score function di p_σ è:

$$\nabla_{\tilde{x}} \log p_\sigma(\tilde{x}) = \nabla_{\tilde{x}} \log \int p_{data}(x)\, q_\sigma(\tilde{x} \mid x)\, dx$$

$$= \left[\int p_{data}(x)\, \nabla_{\tilde{x}}\, q_\sigma(\tilde{x} \mid x)\, dx\right] / \, p_\sigma(\tilde{x})$$

Per $q_\sigma(\tilde{x} \mid x) = (2\pi\sigma^2)^{-n/2} \exp(-\|\tilde{x}-x\|^2/(2\sigma^2))$, il gradiente è:

$$\nabla_{\tilde{x}}\, q_\sigma(\tilde{x} \mid x) = -(\tilde{x}-x)/\sigma^2 \cdot q_\sigma(\tilde{x} \mid x)$$

Quindi:

$$\nabla_{\tilde{x}} \log p_\sigma(\tilde{x}) = -\int (\tilde{x}-x)/\sigma^2 \cdot p_{data}(x)\, q_\sigma(\tilde{x} \mid x)\, dx \, / \, p_\sigma(\tilde{x})$$

$$= -(\tilde{x} - \mathbb{E}[x \mid \tilde{x}]) \, / \, \sigma^2$$

dove $\mathbb{E}[x \mid \tilde{x}] = \int x \cdot p(x \mid \tilde{x})\, dx$ è la media a posteriori di x dato $\tilde{x}$ sotto il modello $p_{data} \times q_\sigma$. Questa è la formula che collega la score function al denoiser ottimale in senso quadratico.

L'obiettivo di denoising score matching con livello di rumore σ è:

$$J_{DSM}(\theta, \sigma) = \tfrac{1}{2}\, \mathbb{E}_{x \sim p_{data},\, \tilde{x} \sim q_\sigma(\cdot \mid x)}\left[\|s_\theta(\tilde{x}) + (\tilde{x}-x)/\sigma^2\|^2\right]$$

Espandendo:

Silvano Salvador

$$= \tfrac{1}{2}\, \mathbb{E}_{p_{data},\, q_\sigma}[\|s_\theta(\tilde{x})\|^2 + 2\, s_\theta(\tilde{x})^T(\tilde{x}-x)/\sigma^2 + \|(\tilde{x}-x)/\sigma^2\|^2]$$

Il terzo termine è costante rispetto a θ. Per il secondo termine:

$$\mathbb{E}_{p_{data},\, q_\sigma}[s_\theta(\tilde{x})^T(\tilde{x}-x)/\sigma^2]$$

$$= \iint s_\theta(\tilde{x})^T(\tilde{x}-x)/\sigma^2 \cdot p_{data}(x)\, q_\sigma(\tilde{x}\,|\,x)\, d\tilde{x}\, dx$$

$$= -\int s_\theta(\tilde{x})^T\, \nabla_{\tilde{x}} \log p_\sigma(\tilde{x}) \cdot p_\sigma(\tilde{x})\, d\tilde{x}$$

$$= -\mathbb{E}_{p_\sigma}[s_\theta(\tilde{x})^T\, \nabla_{\tilde{x}} \log p_\sigma(\tilde{x})]$$

Sostituendo nell'obiettivo di score matching implicito su p_σ:

$$J_{SM}(\theta; p_\sigma) = \mathbb{E}_{p_\sigma}[\mathrm{tr}(\nabla_{\tilde{x}}\, s_\theta(\tilde{x})) + \tfrac{1}{2}\, \|s_\theta(\tilde{x})\|^2]$$

$$= -\mathbb{E}_{p_\sigma}[s_\theta(\tilde{x})^T\, s_{p_\sigma}(\tilde{x})] + \tfrac{1}{2}\, \mathbb{E}_{p_\sigma}[\|s_\theta(\tilde{x})\|^2] + \text{costante}$$

Confrontando con $J_{DSM}(\theta, \sigma)$ e usando l'uguaglianza dimostrata per il secondo termine, si conclude che $J_{DSM}(\theta, \sigma) = J_{SM}(\theta; p_\sigma) +$ costante. Le due funzioni obiettivo hanno quindi lo stesso minimizzatore, e addestrare a denoising equivale formalmente a stimare la score function della distribuzione smoothed.

A.IV.6 — Score matching con perturbazioni multiple e la connessione con NCSN

Song e Ermon (2019) osservarono che stimare la score function su dati non perturbati è difficile nelle regioni a bassa densità della distribuzione, dove i dati di addestramento sono scarsi e la stima diventa imprecisa. La soluzione proposta — i Noise Conditional Score Networks (NCSN) — consiste nell'addestrare un unico modello condizionale $s_\theta(\tilde{x}, \sigma)$ a stimare la score function per un insieme discreto di livelli di rumore $\sigma_1 > \sigma_2 > \ldots > \sigma_L$:

$$J_{NCSN}(\theta) = \sum_{i=1}^{L} \lambda(\sigma_i) \cdot \mathbb{E}_{x,\, \tilde{x}}[\|\sigma_i\, s_\theta(\tilde{x}, \sigma_i) + (\tilde{x}-x)/\sigma_i\|^2]$$

Silvano Salvador

dove $\lambda(\sigma_i)$ è un peso che bilancia i contributi dei diversi livelli di rumore. La scelta $\lambda(\sigma) = \sigma^2$ garantisce che l'obiettivo sia approssimativamente uniforme su tutti i livelli: poiché $\|(\tilde{x}-x)/\sigma\|^2 \sim n$ a ogni livello (per un rumore isotropo con n dimensioni), il prodotto $\sigma^2 \cdot J_DSM$ normalizza il contributo di ogni livello a una scala comparabile.

La generazione con NCSN avviene mediante la dinamica di Langevin annealed: si parte da un campione di rumore puro $z \sim \mathcal{N}(0, \sigma_1^2 I)$ e si simula la catena di Langevin ad ogni livello di rumore, passando progressivamente dal livello più rumoroso σ_1 a quello meno rumoroso σ_L:

Per il livello i, per k = 1, …, K:

$$x_\{k+1\} = x_k + \alpha_i \, s_\theta(x_k, \sigma_i) + \sqrt{(2\alpha_i)}\, z_k, \quad z_k \sim \mathcal{N}(0, I)$$

dove $\alpha_i = \varepsilon \cdot \sigma_i^2/\sigma_L^2$ è il passo di Langevin scalato per il livello di rumore. Al termine dei K passi al livello i, si usa il campione finale come punto di partenza per il livello i+1.

A.IV.7 — DDPM con continuo di livelli di rumore: la formulazione SDE unificata

Song e colleghi (2021) unificano NCSN e DDPM nel linguaggio delle SDE mostrando che entrambi corrispondono a processi di diffusione con strutture di drift e diffusività specifiche. La chiave è identificare il noise schedule $\{\beta_t\}_\{t \in [0,T]\}$ come la funzione di diffusività $g(t)^2 = \beta(t)$ nella SDE:

$$dX_t = -\beta(t)/2 \cdot X_t \, dt + \sqrt{\beta(t)}\, dW_t$$

Questa è la SDE di Ornstein-Uhlenbeck vettoriale. Il marginal $q(x_t \mid x_0)$ si calcola risolvendo la Fokker-Planck associata con condizione iniziale $\delta_\{x_0\}$, ottenendo:

Silvano Salvador

$q(x_t \mid x_0) = \mathcal{N}(x_t; \mu(t)\, x_0, \sigma^2(t)\, I)$

dove le funzioni di mean e varianza soddisfano le ODE:

$d\mu(t)/dt = -\beta(t)/2 \cdot \mu(t), \quad \mu(0) = 1$

$d\sigma^2(t)/dt = \beta(t)(1 - \sigma^2(t)), \quad \sigma^2(0) = 0$

con soluzione:

$\mu(t) = \exp(-\tfrac{1}{2} \int_0^t \beta(s)\, ds) = \sqrt{\bar{\alpha}_t}$

$\sigma^2(t) = 1 - \exp(-\int_0^t \beta(s)\, ds) = 1 - \bar{\alpha}_t$

In coerenza con la derivazione discreta. Per T sufficientemente grande (o $\beta(t)$ crescente abbastanza rapidamente), $\mu(T) \approx 0$ e $\sigma^2(T) \approx 1$, ovvero $q(x_T \mid x_0) \approx \mathcal{N}(0, I)$ indipendentemente da x_0: il processo forward ha effettuato la transizione dalla distribuzione dei dati alla gaussiana standard.

La Probability Flow ODE corrispondente — la versione deterministica del processo reverse — è:

$dx/dt = -\beta(t)/2 \cdot x - \tfrac{1}{2}\, \beta(t)\, \nabla_x \log p_t(x)$

che, sostituendo la stima neurale $s_\theta(x, t) \approx \nabla_x \log p_t(x)$, diventa:

$dx/dt = -\beta(t)/2 \cdot x - \tfrac{1}{2}\, \beta(t)\, s_\theta(x, t) = -\beta(t)/2 \cdot (x + s_\theta(x, t))$

Questa ODE può essere integrata con solvers numerici di ordine elevato — Dormand-Prince, Adams-Moulton — che adattano il passo di integrazione alla curvatura locale della traiettoria. Per traiettorie sufficientemente regolari, questi solvers raggiungono l'accuratezza desiderata con un numero di passi molto inferiore rispetto alla simulazione esplicita della SDE stocastica, spiegando il miglioramento in efficienza dei metodi DDIM rispetto a DDPM originale.

Silvano Salvador

A.IV.8 — Analisi della convergenza: bound teorici sulla qualità della generazione

La qualità dei campioni generati da un modello di diffusione dipende da tre fonti di errore: l'errore di stima della score function — quanto bene s_θ approssima $\nabla_x \log p_t$ — l'errore di discretizzazione — quanto bene la simulazione numerica della SDE reverse approssima il processo continuo — e l'errore di approssimazione del marginal finale — quanto bene $q(x_T \mid x_0)$ approssima $\mathcal{N}(0, I)$.

Benton, De Bortoli, Doucet e Durkan (2023) hanno derivato bound non asintotici sulla divergenza KL tra la distribuzione generata e la distribuzione dei dati, in funzione di questi tre termini di errore. Il risultato principale afferma che, sotto condizioni di regolarità sulla distribuzione dei dati e sulla score function:

$$D_{KL}(p_{gen} \,\|\, p_{data}) \leq C_1 \cdot \varepsilon^2_{score} \cdot T + C_2 \cdot h + C_3 \cdot e^{-\lambda T}$$

dove $\varepsilon^2_{score} = \int_0^T \mathbb{E}_{p_t}[\|s_\theta(x,t) - \nabla_x \log p_t(x)\|^2]\, dt$ è l'errore quadratico integrato della stima della score function, h è il passo di discretizzazione, λ è il tasso di convergenza del processo forward alla gaussiana standard (proporzionale a $\min_{t \in [0,T]} \beta(t)$), e C_1, C_2, C_3 sono costanti che dipendono dalle proprietà della distribuzione dei dati.

Questo bound è importante per diverse ragioni. Primo, esso mostra che la qualità della generazione migliora al migliorare della stima della score function con una dipendenza lineare in T: orizzonte temporale più lungo permette una stima più accurata, ma aumenta anche il costo computazionale. Secondo, esso mostra che l'errore di discretizzazione decresce linearmente nel passo h, giustificando l'uso di solvers di ordine elevato che riducono h per un budget computazionale fisso. Terzo, il

Silvano Salvador

termine esponenziale $e^{-\lambda T}$ quantifica il costo di un orizzonte temporale insufficiente: se T è troppo piccolo, la distribuzione del rumore finale non è ancora gaussiana standard e i campioni generati sono distorti.

Il bound suggerisce che l'investimento computazionale ottimale per un budget fisso debba bilanciare il tempo di addestramento — che riduce $\varepsilon^2_{\text{score}}$ — e il numero di passi di generazione — che riduce l'errore di discretizzazione. Questa analisi del compromesso non è ancora completamente sviluppata nella letteratura, ma costituisce un programma di ricerca attivo che porta verso una teoria quantitativa della qualità generativa dei modelli di diffusione.

A.IV.9 — Generazione condizionata: classifier guidance e classifier-free guidance

Il problema della generazione condizionata — campionare da $p(x \mid c)$ per una condizione c — si inserisce nel framework della diffusione modificando la score function del processo reverse. Poiché:

$$\nabla_x \log p_t(x \mid c) = \nabla_x \log p_t(x) + \nabla_x \log p_t(c \mid x)$$

la score function condizionata è la somma della score function incondizionata e del gradiente del log-classificatore. Nella generazione con classifier guidance, si addestra separatamente un classificatore $p_\varphi(c \mid x_t)$ sulle versioni rumorose dei dati e si usa il gradiente del log-classificatore per guidare la generazione:

$$\nabla_x \log \tilde{p}_t(x \mid c) = s_\theta(x, t) + \gamma \, \nabla_x \log p_\varphi(c \mid x_t)$$

dove $\gamma \geq 1$ è il parametro di intensità del condizionamento. Per $\gamma = 1$ si ottiene la score function bayesiana esatta; per $\gamma > 1$ si amplifica il segnale di condizionamento, spingendo la generazione verso regioni di alta compatibilità con c a costo di una riduzione della diversità.

Silvano Salvador

Il classifier-free guidance di Ho e Salimans (2022) evita l'addestramento di un classificatore separato addestrando un unico modello a stimare sia la score condizionata sia quella incondizionata, impostando c = ∅ con probabilità p_uncond durante l'addestramento. La formula di guidance diventa:

$$\tilde{s}_\theta(x, t, c) = s_\theta(x, t, \varnothing) + \gamma \cdot (s_\theta(x, t, c) - s_\theta(x, t, \varnothing))$$

$$= (1 - \gamma)\, s_\theta(x, t, \varnothing) + \gamma\, s_\theta(x, t, c)$$

Per $\gamma = 1$ si recupera la score condizionata pura; per $\gamma > 1$ si effettua un'estrapolazione lineare nella direzione che allontana dalla distribuzione incondizionata verso quella condizionata. L'interpretazione geometrica è che si campiona non dalla distribuzione p_t(x | c) ma da una distribuzione modificata con densità proporzionale a p_t(x | c)^γ / p_t(x)^{γ−1}, che è più concentrata sulle regioni di alta compatibilità con c rispetto alla distribuzione condizionata standard.

La superiorità empirica del classifier-free guidance rispetto al classifier guidance nelle applicazioni text-to-image — documentata nei lavori su DALL-E 2, Imagen e Stable Diffusion — si spiega con due fattori: l'eliminazione del disallineamento tra il classificatore addestrato su dati rumorosi e il modello di diffusione, e la capacità di modulare l'intensità del condizionamento senza richiedere la ricalibrazione di un classificatore separato per ogni valore di γ.

Silvano Salvador

Appendice Tecnica V

Normalizing flows: dimostrazioni complete, architetture invertibili e geometria dello spazio delle distribuzioni

A.V.1 — Il teorema del cambiamento di variabile: fondamenta

Il teorema del cambiamento di variabile per variabili casuali continue è il risultato analitico che rende possibile l'intera architettura dei normalizing flows. Nella sua forma generale, esso afferma che se X è una variabile casuale continua con densità $p_X(x)$ e $f : \mathbb{R}^n \to \mathbb{R}^n$ è una funzione differenziabile e invertibile (un diffeomorfismo), allora la variabile trasformata $Y = f(X)$ ha densità:

$$p_Y(y) = p_X(f^{-1}(y)) \cdot |\det J_{f^{-1}}(y)|$$

$$= p_X(f^{-1}(y)) \,/\, |\det J_f(f^{-1}(y))|$$

dove $J_f(x)$ è la matrice Jacobiana di f valutata in x, con elementi $[J_f(x)]_{ij} = \partial f_i/\partial x_j$. La seconda uguaglianza usa il teorema della funzione inversa: $J_{f^{-1}}(y) = [J_f(f^{-1}(y))]^{-1}$ e $|\det A^{-1}| = 1/|\det A|$.

La derivazione procede per il teorema di trasformazione delle misure. La misura di probabilità di Y nell'insieme $B \subseteq \mathbb{R}^n$ è:

$$P(Y \in B) = P(X \in f^{-1}(B)) = \int_{f^{-1}(B)} p_X(x)\, dx$$

Applicando il teorema del cambiamento di variabile per integrali multipli con la sostituzione $x = f^{-1}(y)$:

$$= \int_B p_X(f^{-1}(y))\, |\det J_{f^{-1}}(y)|\, dy$$

Per la validità dell'integrale, si richiede che f sia un diffeomorfismo C^1 — continuamente differenziabile con jacobiano invertibile ovunque —

Silvano Salvador

condizione che garantisce che la sostituzione di variabile sia legittima e che la densità risultante sia non negativa e si integri a uno.

La formula per log p_Y si ottiene prendendo il logaritmo:

$$\log p_Y(y) = \log p_X(f^{-1}(y)) + \log \left|\det J_{f^{-1}}(y)\right|$$

$$= \log p_X(f^{-1}(y)) - \log \left|\det J_f(f^{-1}(y))\right|$$

In un normalizing flow a K strati, con trasformazioni f_1, f_2, …, f_K applicate sequenzialmente a partire da $z_0 \sim p_z$:

$$z_k = f_k(z_{k-1}), \quad k = 1, \ldots, K$$

$$x = z_K$$

la log-densità di x si calcola per applicazione iterativa della formula:

$$\log p(x) = \log p_z(z_0) - \sum_{k=1}^{K} \log \left|\det J_{f_k}(z_{k-1})\right|$$

dove $z_0 = f_K^{-1} \circ f_{K-1}^{-1} \circ \ldots \circ f_1^{-1}(x)$ è ottenuto invertendo la composizione. La log-verosimiglianza è dunque esattamente calcolabile — senza approssimazioni variazionali — per qualsiasi input x, purché si disponga delle trasformazioni inverse e del determinante jacobiano a ogni strato.

A.V.2 — La scelta della base p_z e la geometria della trasformazione

La distribuzione base p_z deve essere semplice da campionare e da valutare. La scelta quasi universale è la gaussiana standard $\mathcal{N}(0, I_d)$: essa è log-concava, isotropa, e la sua log-densità

$$\log p_z(z) = -d/2 \cdot \log(2\pi) - \tfrac{1}{2} \|z\|^2$$

è calcolabile in O(d). La scelta della distribuzione base determina la geometria dello spazio latente: poiché le trasformazioni f_k sono

Silvano Salvador

diffeomorfismi, la topologia della distribuzione generata è la stessa della distribuzione base. Una gaussiana standard — topologicamente omeomorfa a $\mathbb{R}^n$ — non può essere trasformata da un diffeomorfismo globale in una distribuzione con più componenti connesse separate, come avverrebbe per una distribuzione bimodale con modo ben separati. Questo è un limite topologico strutturale dei normalizing flows con distribuzione base connessa.

La questione della capacità espressiva dei normalizing flows è legata alla teoria dei diffeomorfismi. Il teorema di Brenier (1991) afferma che per ogni coppia di misure P e Q con stessa massa totale e finite aspettative, esiste un unico mappa di trasporto ottimale T* che trasforma P in Q minimizzando il costo quadratico $\int \|x - T(x)\|^2 \, dP(x)$. Questa mappa è il gradiente di una funzione convessa — $T^*(x) = \nabla \varphi(x)$ per una funzione φ convessa — ed è la soluzione del problema di Monge con costo quadratico. Il flow ottimale che realizza il trasporto di Monge-Kantorovich non è in generale realizzabile da un network di coupling layers finite, il che implica che qualsiasi architettura di normalizing flow introduce un errore di approssimazione rispetto alla classe ottimale di trasformazioni.

A.V.3 — NICE e RealNVP: analisi completa delle coupling layers

Le coupling layers di Dinh e colleghi (2014, 2016) costruiscono trasformazioni invertibili con jacobiano triangolare a blocchi, il cui determinante è calcolabile in O(d). La struttura fondamentale per RealNVP è la seguente.

Dato l'input $x \in \mathbb{R}^d$, si partiziona x in due parti: x = (x_A, x_B) dove x_A $\in \mathbb{R}$^{d_A} e x_B $\in \mathbb{R}$^{d_B} con d_A + d_B = d. La coupling layer produce:

Silvano Salvador

$y_A = x_A$

$y_B = x_B \odot \exp(s(x_A)) + t(x_A)$

dove $s, t : \mathbb{R}^{d_A} \to \mathbb{R}^{d_B}$ sono reti neurali arbitrarie (i funzionali di scala e traslazione), e $\odot$ denota il prodotto di Hadamard componente per componente. L'output è $y = (y_A, y_B) \in \mathbb{R}^d$.

L'invertibilità è immediata per costruzione: da $y = (y_A, y_B)$ si recupera $x_A = y_A$, e poi $x_B = (y_B - t(y_A)) \odot \exp(-s(y_A))$. Notare crucialmente che le reti s e t non devono essere invertibili: vengono valutate soltanto sulla parte A dell'input, che rimane invariata, per cui non compare nessuna equazione implicita nell'inversione.

La matrice jacobiana $\partial y/\partial x$ ha struttura a blocchi:

$$J = \partial(y_A, y_B)/\partial(x_A, x_B) = \begin{bmatrix} I_{d_A} & 0 \\ \partial y_B/\partial x_A & \operatorname{diag}(\exp(s(x_A))) \end{bmatrix}$$

La sottomatrice $\partial y_B/\partial x_A$ contiene le derivate delle funzioni s e t rispetto a x_A, ed è in generale densa. La sottomatrice $\partial y_B/\partial x_B = \operatorname{diag}(\exp(s(x_A)))$ è diagonale perché ogni componente $y_{B,i}$ dipende da $x_{B,i}$ soltanto attraverso il termine additivo $x_{B,i} \cdot \exp(s_i(x_A))$.

Poiché J è triangolare a blocchi — il blocco superiore destro è zero — il determinante è il prodotto dei determinanti dei blocchi diagonali:

$$\det J = \det(I_{d_A}) \cdot \det(\operatorname{diag}(\exp(s(x_A))))$$

$$= 1 \cdot \prod_i \exp(s_i(x_A))$$

$$= \exp\left(\sum_i s_i(x_A)\right)$$

Quindi:

$$\log |\det J| = \sum_i s_i(x_A) = 1^T s(x_A)$$

Silvano Salvador

un'operazione O(d) indipendente dalla complessità delle reti s e t. Questo è il risultato chiave: per quanto complesse siano le reti s e t — qualsiasi architettura, qualsiasi numero di parametri — il log-determinante è semplicemente la somma degli output del funzionale di scala.

Il limite della coupling layer è che la parte A dell'input non viene mai trasformata direttamente: y_A = x_A identicamente. Per garantire che ogni componente venga trasformata, RealNVP alterna coupling layers con diverse partizioni di (A, B): per esempio, si usa la parità dell'indice come criterio di partizionamento in modo alternato. Dopo K strati con partizioni alternate, ogni componente è stata trasformata in almeno K/2 strati. Dinh e colleghi (2016) usano anche permutazioni casuali delle componenti tra gli strati per aumentare il mixing.

A.V.4 — Autoregressive flows: MAF e IAF

Le autoregressive flows sfruttano la fattorizzazione della densità come prodotto di condizionali:

$$p(x) = \prod_{i=1}^{d} p(x_i \mid x_{<i}) = \prod_{i=1}^{d} p(x_i \mid x_1, \ldots, x_{i-1})$$

parametrizzando ogni condizionale come una trasformazione invertibile scalare di una variabile di rumore:

$$x_i = g(\varepsilon_i; h_i(x_{<i})), \quad \varepsilon_i \sim \mathcal{N}(0, 1)$$

dove $h_i : \mathbb{R}^{i-1} \to \mathbb{R}^{k}$ sono i parametri della trasformazione condizionale e g è una funzione monotona (tipicamente una traslazione scalata: $x_i = \mu_i(x_{<i}) + \sigma_i(x_{<i}) \cdot \varepsilon_i$ per il caso gaussiano).

La jacobiana di $x = g(\varepsilon; h(x_{<i}))$ rispetto a ε è diagonale con elementi $\sigma_i(x_{<i})$, il che produce:

$$\log |\det J_{x/\varepsilon}| = \sum_i \log \sigma_i(x_{<i})$$

Silvano Salvador

La struttura autoregressiva garantisce che la jacobiana della trasformazione $\varepsilon \mapsto x$ sia triangolare inferiore: $\partial x_i / \partial \varepsilon_j = 0$ per $j > i$ poiché x_i non dipende da ε_j con $j > i$. Il log-determinante è dunque la somma delle log-deviazioni standard condizionali, calcolabile in O(d).

Il Masked Autoregressive Flow (MAF) di Papamakarios, Pavlakou e Murray (2017) implementa le dipendenze autoreggressive usando MADE (Masked Autoencoder for Distribution Estimation), una rete neurale con mascheratura dei pesi che garantisce la struttura autoregressiva con un singolo forward pass. Nell'architettura MADE, ogni connessione dalla dimensione j all'unità nascosta k è attiva soltanto se il grado assegnato a j è strettamente inferiore al grado assegnato a k, e ogni connessione dall'unità nascosta k alla dimensione di output i è attiva soltanto se il grado di k è inferiore a i. Questa mascheratura garantisce che l'output i dipenda soltanto dagli input $j < i$, implementando la struttura autoregressiva senza calcoli sequenziali durante il forward pass.

Il limite di MAF è che la generazione di nuovi campioni è sequenziale: per campionare x, si deve calcolare $x_1 = g(\varepsilon_1; h_1(\varnothing))$, poi $x_2 = g(\varepsilon_2; h_2(x_1))$, poi $x_3 = g(\varepsilon_3; h_3(x_1, x_2))$, e così via per d passi. Questo rende la generazione O(d) non parallelizzabile.

L'Inverse Autoregressive Flow (IAF) di Kingma e colleghi (2016) inverte la direzione della trasformazione: parametrizza il flusso nella direzione $\varepsilon \mapsto x$, con $x_i = \mu_i(\varepsilon_{<i}) + \sigma_i(\varepsilon_{<i}) \cdot \varepsilon_i$. In questa direzione la generazione è parallelizzabile — tutte le componenti x_i possono essere calcolate simultaneamente dato ε, poiché h_i dipende da $\varepsilon_{<i}$ che è già noto — mentre la valutazione della densità $p(x)$ è sequenziale, poiché richiede l'inversione della trasformazione per recuperare ε da x. MAF e IAF sono duali in questo senso: MAF è efficiente per la stima della

Silvano Salvador

densità (addestramento) e lento per la generazione; IAF è lento per la stima della densità e veloce per la generazione.

A.V.5 — Glow: convoluzioni 1×1 invertibili e decomposizione LU

Kingma e Dhariwal (2018) introducono in Glow due contributi tecnici principali: le convoluzioni 1×1 invertibili come generalizzazione delle permutazioni tra coupling layers, e la parametrizzazione delle convoluzioni invertibili tramite decomposizione LU per rendere il calcolo del determinante O(d) anziché $O(d^3)$.

Una convoluzione 1×1 su un tensore di feature $h \in \mathbb{R}^{H\times W\times C}$ applica una matrice $W \in \mathbb{R}^{C\times C}$ a ogni posizione spaziale indipendentemente:

$y_{h,w} = W x_{h,w}$ per ogni (h,w)

Il log-determinante della jacobiana totale (su tutte le H·W posizioni) è:

$\log |\det J| = H \cdot W \cdot \log |\det W|$

Il problema è che il calcolo di det W ha costo $O(C^3)$ nella dimensione dei canali, che per reti profonde con C = 512 o C = 1024 diventa proibitivo.

La soluzione di Glow è parametrizzare W tramite decomposizione PLU:

$W = P L U$

dove P è una matrice di permutazione (fissa, scelta casualmente e poi tenuta costante), L è una matrice triangolare inferiore con diagonale unitaria, e U è una matrice triangolare superiore. Il determinante di W è:

$\det W = \det P \cdot \det L \cdot \det U = \pm 1 \cdot 1 \cdot \prod_i U_{ii}$

Silvano Salvador

poiché det P = ±1 per qualsiasi permutazione, det L = 1 per una matrice triangolare con diagonale unitaria, e det U è il prodotto degli elementi diagonali. Quindi:

log |det W| = ∑ᵢ log |U_{ii}|

un'operazione O(C). I parametri da apprendere sono gli elementi di L sotto la diagonale (C(C−1)/2 parametri) e gli elementi di U inclusa la diagonale (C(C+1)/2 parametri), per un totale di C^2 parametri — lo stesso del caso generale, ma con il determinante calcolabile in O(C).

L'inversione di W si calcola come $U^{-1}\ L^{-1}\ P^T$, dove le inversioni di U e L sono $O(C^2)$ per matrici triangolari, anzichè $O(C^3)$ per matrici generali.

A.V.6 — Continuous normalizing flows e l'equazione dell'istantaneous change of variables

I Continuous Normalizing Flows (CNF), introdotti da Chen e colleghi (2018) contestualmente all'articolo sulle Neural ODEs, modellano il flow come un processo continuo governato da un'ODE:

dz(t)/dt = f(z(t), t; θ), $z(0) = z_0 \sim p_0$

dove f è un campo vettoriale parametrizzato da una rete neurale. La densità di z(t) evolve secondo l'equazione del cambiamento di variabile istantaneo:

d log p(z(t))/dt = −tr(∂f/∂z(t))

Questa è una conseguenza della formula di Liouville — che afferma che per flussi divergence-free il volume si conserva — nella forma per densità di probabilità. La derivazione procede dall'equazione di continuità per la densità:

∂_t p(z, t) = −div(p(z, t) f(z, t)) = −∇_z p · f − p · tr(∂f/∂z)

Silvano Salvador

Dividendo per p(z, t) e riconoscendo che d log p(z(t))/dt = (1/p) · (∂_t p + f · ∇_z p):

$$\frac{d \log p(z(t))}{dt} = -\mathrm{tr}\left(\frac{\partial f}{\partial z}\right)$$

Integrando da t=0 a t=T:

$$\log p(z(T)) = \log p(z(0)) - \int_0^T \mathrm{tr}\left(\frac{\partial f}{\partial z(t)}\right) dt$$

La log-densità di z(T) è dunque la log-densità iniziale meno l'integrale della traccia della jacobiana del campo vettoriale lungo la traiettoria. Questo elimina la necessità di invertire la trasformazione per valutare la densità: la stessa ODE usata per propagare z forward produce anche la log-densità attraverso il termine aggiuntivo tr(∂f/∂z).

Il costo computazionale dell'approccio è il calcolo della traccia $\int_0^T \mathrm{tr}(\partial f/\partial z)\, dt$, che per un campo vettoriale generico richiederebbe $O(d^2)$ operazioni per ogni passo di integrazione. L'estimatore di Hutchinson riduce questo a O(d) con stima stocastica della traccia, usando un vettore casuale $\varepsilon \sim \mathcal{N}(0, I)$:

$$\mathrm{tr}\left(\frac{\partial f}{\partial z}\right) \approx \varepsilon^T \left(\frac{\partial f}{\partial z}\right) \varepsilon = \varepsilon^T \frac{\partial (f^T \varepsilon)}{\partial z}$$

dove l'ultimo termine è il prodotto scalare tra il vettore ε e il gradiente del prodotto scalare $f^T\varepsilon$ rispetto a z — calcolabile con un singolo backward pass di autodifferenziazione.

A.V.7 — Flow Matching: addestramento senza simulare l'ODE

Lipman e colleghi (2022) e Liu e colleghi (2022) introducono il Flow Matching come metodo per addestrare continuous normalizing flows senza simulare l'ODE durante l'addestramento — una limitazione significativa del training dei CNF originali, dove ogni gradiente richiedeva la risoluzione numerica dell'ODE forward e backward.

Silvano Salvador

L'idea chiave è di specificare direttamente il campo vettoriale obiettivo che dovrebbe essere appreso, invece di derivarlo minimizzando la log-verosimiglianza. Si definisce un probability path $\{p_t\}_{t \in [0,1]}$ che interpola tra la distribuzione base $p_0 = \mathcal{N}(0, I)$ e la distribuzione dei dati $p_1 = p_{data}$. Per ogni coppia (x_0, x_1) con $x_0 \sim p_0$ e $x_1 \sim p_1$, si definisce un percorso condizionale:

$x_t \mid (x_0, x_1) = (1-t)\, x_0 + t\, x_1$ (interpolazione lineare)

Il campo vettoriale marginale che genera il probability path è:

$u_t(x) = \mathbb{E}[dx_t/dt \mid x_t = x] = \mathbb{E}[x_1 - x_0 \mid x_t = x]$

Questo campo è difficile da calcolare esattamente (richiede la stima della distribuzione condizionale $p(x_0, x_1 \mid x_t = x)$). Il Flow Matching sostituisce l'obiettivo marginale con l'obiettivo condizionale, che è accessibile perché la condizione (x_0, x_1) è nota durante l'addestramento:

$L_{CFM}(\theta) = \mathbb{E}_{t, x_0, x_1}[\, \| v_\theta(x_t, t) - (x_1 - x_0) \|^2]$

dove v_θ è la rete neurale che approssima il campo vettoriale obiettivo. Chen e Lipman (2023) dimostrano che L_{CFM} e l'obiettivo Flow Matching marginale hanno lo stesso gradiente rispetto a θ, per cui minimizzare L_{CFM} su campioni (t, x_0, x_1) produce lo stesso campo vettoriale del Flow Matching. La stima dell'obiettivo condizionale richiede un solo forward pass della rete e nessuna soluzione ODE: l'addestramento è $O(batch_size \cdot d)$, confrontabile con quello dei modelli di diffusione.

La connessione con i DDPM è immediata: il percorso condizionale lineare $x_t = (1-t)x_0 + t\, x_1$ è un caso speciale del processo di diffusione con noise schedule lineare. Il Flow Matching con percorso lineare produce campi vettoriali rettilinei — la rete apprende a muoversi direttamente dal rumore verso i dati — che sono più facili da integrare

Silvano Salvador

numericamente rispetto ai percorsi curvilinei tipici dei modelli di diffusione con noise schedule coseno. Questo si traduce in generazione accurata con meno passi di integrazione dell'ODE, combinando i vantaggi della qualità dei modelli di diffusione con l'efficienza della generazione in pochi passi.

A.V.8 — Espressività universale dei normalizing flows

Una questione teorica fondamentale è se una composizione finita di trasformazioni di un tipo dato — per esempio, coupling layers — sia universalmente espressiva nel senso che può approssimare arbitrariamente bene qualsiasi diffeomorfismo su $\mathbb{R}^n$. Huang e colleghi (2018) hanno dimostrato, per le Neural Autoregressive Flows (NAF), che le autoregressive flows con funzioni di trasformazione monotone parametrizzate da reti neurali arbitrariamente profonde sono universalmente espressive nella classe dei diffeomorfismi. Il risultato usa la densità delle reti neurali nella classe delle funzioni continue (teorema di approssimazione universale di Cybenko e Hornik) combinata con la struttura autoregressiva che garantisce l'invertibilità.

Per le coupling layers, il risultato di espressività universale è più delicato poiché la struttura della partizione in (A, B) limita le dipendenze che uno strato singolo può catturare. Toth e colleghi (2019) hanno dimostrato che una composizione di coupling layers con partizioni alternate è universalmente espressiva per trasformazioni che hanno una struttura di dipendenza compatibile con le partizioni scelte. In pratica, per dati ad alta dimensionalità con struttura di correlazione complessa, può essere necessario un numero molto elevato di strati per approssimare con precisione la distribuzione target — un limite pratico che le architetture ibride come i modelli di diffusione non incontrano nella stessa misura.

Silvano Salvador

A.V.9 — Confronto con i modelli di diffusione: qualità, efficienza e garanzie

Il confronto tra normalizing flows e modelli di diffusione può essere strutturato lungo tre dimensioni: la qualità delle densità stimate, l'efficienza della generazione, e le garanzie teoriche disponibili.

Sul piano della qualità delle densità, i normalizing flows offrono log-verosimiglianza esatta: per ogni dato x di test, log p(x) è calcolabile esattamente come log $p_z(f^{-1}(x)) - \log |\det J_f(f^{-1}(x))|$. Questo rende i flows la scelta naturale per le applicazioni di stima di densità, rilevamento di anomalie, e comparazione quantitativa di modelli generativi. I modelli di diffusione forniscono soltanto un lower bound alla log-verosimiglianza — l'ELBO — il cui gap rispetto alla log-verosimiglianza vera non è direttamente misurabile.

Sul piano dell'efficienza della generazione, la situazione è inversa. I normalizing flows devono applicare la composizione di K trasformazioni a ogni campione generato, dove K è il numero di strati necessario per la qualità desiderata. Per dati ad alta dimensionalità e distribuzione complessa — immagini ad alta risoluzione, audio, video — K deve essere molto elevato, rendendo la generazione computazionalmente costosa. I modelli di diffusione con Flow Matching e solvers ODE adattativi possono generare campioni di alta qualità in 10-50 valutazioni della rete, confrontabili con i flows per dimensione del modello comparabile.

Sul piano delle garanzie teoriche, i flows offrono garanzie di invertibilità e di calcolo esatto del determinante per costruzione, mentre i modelli di diffusione richiedono soltanto la qualità della stima della score function e la precisione del solver numerico. I flows non allucinano nel senso che ogni punto dello spazio dati ha una densità calcolabile — non c'è regione dello spazio con densità indefinita — ma possono produrre campioni di

Silvano Salvador

bassa qualità se la distribuzione appresa non copre adeguatamente la distribuzione reale, esattamente come qualsiasi altro modello generativo.

La convergenza pratica verso architetture ibride — flow matching applicato a spazi latenti di autoencoder, come in Latent Diffusion Models (Rombach et al., 2022) — suggerisce che le due famiglie di modelli non siano in competizione ma complementari: i flows sono ottimali per la stima di densità in spazi di moderata dimensionalità; i modelli di diffusione sono ottimali per la generazione di alta qualità in spazi ad altissima dimensionalità. La distinzione è, ancora una volta, un compromesso tra fedeltà analitica e scalabilità empirica — la tensione fondamentale che percorre l'intera architettura dell'intelligenza artificiale moderna.

Bibliografia critica annotata

Fondamenti della teoria dell'informazione

Shannon, C.E. (1948). A Mathematical Theory of Communication. Bell System Technical Journal, 27, 379-423; 623-656. Fondativo. Introduce la misura di entropia, il teorema di codifica della sorgente e il teorema di codifica del canale. La derivazione assiomatica dell'entropia come unica misura soddisfacente le condizioni di continuità, simmetria e consistenza ricorsiva rimane il risultato definitivo.

Kullback, S., Leibler, R.A. (1951). On Information and Sufficiency. Annals of Mathematical Statistics, 22(1), 79-86. Introduce la divergenza che porta il loro nome come misura di distinguibilità tra distribuzioni e la collega al lemma di Neyman-Pearson per i test di ipotesi.

Silvano Salvador

Jaynes, E.T. (1957). Information Theory and Statistical Mechanics, I e II. Physical Review, 106(4), 620-630; 108(2), 171-190. Rifondazione della meccanica statistica come applicazione del principio di massima entropia. Ancora il più chiaro argomento per l'interpretazione epistemica dell'entropia.

Amari, S. (1985). Differential-Geometrical Methods in Statistics. Springer. Sviluppo sistematico della geometria dell'informazione. La connessione tra metrica di Fisher, divergenze e varietà esponenziali è esposta con completezza formale.

Cover, T.M., Thomas, J.A. (2006). Elements of Information Theory. 2a ed. Wiley. Il riferimento standard per la trattazione moderna della teoria dell'informazione. Capitoli 2-5 per i contenuti della Parte Prima.

Inferenza bayesiana e modelli variazionali

Cox, R.T. (1946). Probability, Frequency, and Reasonable Expectation. American Journal of Physics, 14(1), 1-13. Fondativo per la tradizione bayesiana soggettivista. Dimostra che le uniche misure di credenza coerenti con la logica proposizionale sono le misure di probabilità.

Jordan, M.I., Ghahramani, Z., Jaakkola, T., Saul, L.K. (1999). An Introduction to Variational Methods for Graphical Models. Machine Learning, 37(2), 183-233. Introduzione sistematica all'inferenza variazionale nei modelli grafici.

Kingma, D.P., Welling, M. (2014). Auto-Encoding Variational Bayes. ICLR 2014. Introduce il VAE e il trucco della riparametrizzazione. Il paper originale è notevolmente compatto: tre pagine di derivazione per un'architettura che ha generato migliaia di varianti.

Silvano Salvador

Burda, Y., Grosse, R., Salakhutdinov, R. (2016). Importance Weighted Autoencoders. ICLR 2016. Introduce l'IWAE come bound più stretto dell'ELBO mediante campionamento per importanza.

Kleijn, B.J.K., van der Vaart, A.W. (2012). The Bernstein-Von-Mises Theorem Under Misspecification. Electronic Journal of Statistics, 6, 354-381. Risultato cruciale per la comprensione dei limiti dell'inferenza bayesiana. Dimostra la convergenza della distribuzione a posteriori attorno alla pseudo-verità con covarianza sandwich sotto misspecification.

Higgins, I., Matthey, L., Pal, A., et al. (2017). beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. ICLR 2017. Introduce la penalizzazione variabile del termine KL per ottenere rappresentazioni disentangled.

Finn, C., Abbeel, P., Levine, S. (2017). Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. ICML 2017. Il paper fondativo del meta-apprendimento basato su gradiente, interpretabile come apprendimento di prior impliciti.

Termodinamica dell'informazione e modelli di diffusione

Landauer, R. (1961). Irreversibility and Heat Generation in the Computing Process. IBM Journal of Research and Development, 5(3), 183-191. Stabilisce il collegamento quantitativo tra cancellazione dell'informazione e dissipazione di energia.

Bennett, C.H. (2003). Notes on Landauer's Principle, Reversible Computation, and Maxwell's Demon. Studies in History and Philosophy of Science Part B, 34(3), 501-510. Chiarificazione del principio di Landauer: la misurazione è reversibile, la cancellazione no.

Silvano Salvador

Hopfield, J.J. (1982). Neural Networks and Physical Systems with Emergent Collective Computational Abilities. Proceedings of the National Academy of Sciences, 79(8), 2554-2558. Introduce le reti neurali come sistemi fisici con funzione di energia. Il collegamento tra memoria e minimizzazione dell'energia è stabilito con eleganza sorprendente per un articolo di sei pagine.

Ramsauer, H., et al. (2020). Hopfield Networks is All You Need. ICLR 2021. Dimostra che le Modern Hopfield Networks con energia esponenziale hanno capacità di memorizzazione esponenziale e che il meccanismo di attenzione dei transformer è un singolo passo di aggiornamento di una Hopfield network.

Hyvärinen, A. (2005). Estimation of Non-Normalized Statistical Models by Score Matching. Journal of Machine Learning Research, 6, 695-709. Introduce lo score matching implicito come alternativa alla massimizzazione della verosimiglianza per modelli con funzione di partizione intrattabile.

Vincent, P. (2011). A Connection Between Score Matching and Denoising Autoencoders. Neural Computation, 23(7), 1661-1674. Dimostra l'equivalenza tra denoising score matching e score matching implicito sulla distribuzione smoothed.

Song, Y., Sohl-Dickstein, J., Kingma, D.P., et al. (2021). Score-Based Generative Modeling through Stochastic Differential Equations. ICLR 2021. Unifica DDPM e NCSN nel linguaggio delle SDE, introduce la Probability Flow ODE e la SDE reverse di Anderson.

Ho, J., Jain, A., Abbeel, P. (2020). Denoising Diffusion Probabilistic Models. NeurIPS 2020. Il paper che ha reso pratici i modelli di diffusione, introducendo la perdita L_simple e il noise schedule discreta.

Silvano Salvador

Anderson, B.D.O. (1982). Reverse-time Diffusion Equation Models. Stochastic Processes and their Applications, 12(3), 313-326. Il risultato teorico fondamentale sulla reversibilità dei processi di diffusione, alla base di tutti i modelli generativi di diffusione.

Friston, K. (2010). The Free-Energy Principle: A Unified Brain Theory? Nature Reviews Neuroscience, 11(2), 127-138. L'esposizione più accessibile del principio di energia libera variazionale applicato alla teoria del cervello.

Colombo, M., Seriès, P. (2012). Bayes in the Brain — On Bayesian Modelling in Neuroscience. British Journal for the Philosophy of Science, 63(3), 697-723. Critica metodologica rigorosa al framework di Friston. Argomenta che il principio di energia libera nella sua formulazione generale è non falsificabile.

Rao, R.P.N., Ballard, D.H. (1999). Predictive Coding in the Visual Cortex: A Functional Interpretation of Some Extra-classical Receptive-Field Effects. Nature Neuroscience, 2(1), 79-87. Formulazione del predictive coding come implementazione dell'inferenza variazionale nelle gerarchie corticali.

Compressione, leggi di scala e sovreparametrizzazione

Rissanen, J. (1996). Fisher Information and Stochastic Complexity. IEEE Transactions on Information Theory, 42(1), 40-47. Derivazione del codice NML e della complessità parametrica. Il collegamento tra MDL, misura di Jeffreys e volume riemanniano dello spazio dei parametri.

Solomonoff, R.J. (1964). A Formal Theory of Inductive Inference. Information and Control, 7(1-2), 1-22; 224-254. Introduzione del prior universale e del predittore di Solomonoff. Il risultato di dominanza universale è qui formulato per la prima volta.

Silvano Salvador

Kaplan, J., et al. (2020). Scaling Laws for Neural Language Models. arXiv:2001.08361. Le prime leggi di scala sistematiche per i modelli linguistici, con gli esponenti α_N, α_D, α_C.

Hoffmann, J., et al. (2022). Training Compute-Optimal Large Language Models. NeurIPS 2022. Il lavoro Chinchilla. Raffina le leggi di scala con la formula congiunta L(N,D) e deriva la regola di allocazione ottimale N* ∝ D* ∝ C^0.5.

Belkin, M., Hsu, D., Ma, S., Mandal, S. (2019). Reconciling Modern Machine-Learning Practice and the Classical Bias-Variance Trade-Off. Proceedings of the National Academy of Sciences, 116(32), 15849-15854. Documentazione sistematica della doppia discesa.

Bartlett, P.L., Montanari, A., Rakhlin, A., Zhou, A. (2020). Benign Overfitting in Linear Regression. Proceedings of the National Academy of Sciences, 117(48), 30063-30070. Dimostrazione rigorosa che la soluzione di norma minima generalizza nel regime sovra-parametrizzato.

Tishby, N., Zaslavsky, N. (2015). Deep Learning and the Information Bottleneck Principle. IEEE Information Theory Workshop 2015. Formulazione dell'ipotesi del collo di bottiglia e della fase di compressione durante l'addestramento.

Saxe, A.M., Bansal, Y., Dapello, J., et al. (2018). On the Information Bottleneck Theory of Deep Learning. ICLR 2018. Confutazione della fase di compressione per reti con attivazioni deterministiche.

Wei, J., et al. (2022). Emergent Abilities of Large Language Models. Transactions on Machine Learning Research, 2022. Documentazione delle capacità emergenti oltre soglie dimensionali critiche.

Silvano Salvador

Schaeffer, R., Miranda, B., Koyejo, S. (2023). Are Emergent Abilities of Large Language Models a Mirage? NeurIPS 2023. Mostra che le emergent abilities sono artefatti delle metriche di valutazione discontinue.

Modelli generativi, trasporto ottimale e questioni filosofiche

Goodfellow, I., et al. (2014). Generative Adversarial Nets. NeurIPS 2014. Il paper fondativo delle GAN con la derivazione della divergenza JS come obiettivo ottimale del generatore.

Arjovsky, M., Chintala, S., Bottou, L. (2017). Wasserstein GAN. ICML 2017. Introduce la distanza di Wasserstein-1 come obiettivo per le GAN, risolvendo il problema del gradiente nullo con supporti disgiunti.

Gulrajani, I., Ahmed, F., Arjovsky, M., et al. (2017). Improved Training of Wasserstein GANs. NeurIPS 2017. Introduce la gradient penalty come alternativa al clipping dei pesi per imporre il vincolo di Lipschitz.

Dinh, L., Sohl-Dickstein, J., Bengio, S. (2016). Density Estimation Using Real-Valued Non-Volume Preserving Transformations. ICLR 2017. Il paper RealNVP: coupling layers con log-determinante O(d).

Kingma, D.P., Dhariwal, P. (2018). Glow: Generative Flow with Invertible 1x1 Convolutions. NeurIPS 2018. Introduce le convoluzioni 1×1 invertibili con decomposizione LU per il calcolo efficiente del determinante.

Lipman, Y., Chen, R.T.Q., Ben-Hamu, H., et al. (2022). Flow Matching for Generative Modeling. ICLR 2023. Introduce il Flow Matching come addestramento di CNF senza simulazione dell'ODE.

Levy, O., Goldberg, Y. (2014). Neural Word Embedding as Implicit Matrix Factorization. NeurIPS 2014. Dimostra che word2vec fattorizza implicitamente la matrice di PMI.

Silvano Salvador

Radford, A., Kim, J.W., Hallacy, C., et al. (2021). Learning Transferable Visual Models From Natural Language Supervision. ICML 2021. Il paper CLIP: apprendimento di spazio latente comune testo-immagine con obiettivo contrastivo.

Ho, J., Salimans, T. (2022). Classifier-Free Diffusion Guidance. NeurIPS 2022 Workshop. Introduce il classifier-free guidance, eliminando la necessità di un classificatore separato per la generazione condizionata.

Searle, J.R. (1980). Minds, Brains, and Programs. Behavioral and Brain Sciences, 3(3), 417-424. L'argomento della stanza cinese. Con commenti e risposte nella stessa pubblicazione.

Polanyi, M. (1966). The Tacit Dimension. Doubleday. Il concetto di tacit knowledge come forma di sapere incorporato non riducibile a regole esplicite.

Lake, B.M., Baroni, M. (2018). Generalization without Systematicity: On the Compositional Skills of Sequence-to-Sequence Recurrent Networks. ICML 2018. Il benchmark SCAN. Mostra che le reti seq2seq falliscono sulla generalizzazione composizionale sistematica.

Xie, S.M., Raghunathan, A., Liang, P., Ma, T. (2022). An Explanation of In-Context Learning as Implicit Bayesian Inference. ICLR 2022. Formalizzazione dell'in-context learning come inferenza bayesiana implicita sulla distribuzione del concetto latente.

Teoria dei giochi e allineamento

von Neumann, J., Morgenstern, O. (1944). Theory of Games and Economic Behavior. Princeton University Press. Il testo fondativo. Il teorema minimax per giochi a somma zero è dimostrato nel capitolo 17.

Silvano Salvador

Nash, J.F. (1950). Equilibrium Points in n-Person Games. Proceedings of the National Academy of Sciences, 36(1), 48-49. Due pagine. La dimostrazione dell'esistenza dell'equilibrio in strategie miste come conseguenza del teorema di Kakutani.

Aumann, R.J. (1974). Subjectivity and Correlation in Randomized Strategies. Journal of Mathematical Economics, 1(1), 67-96. Introduce l'equilibrio correlato. Più computazionalmente trattabile dell'equilibrio di Nash e più naturalmente raggiungibile attraverso dinamiche di apprendimento.

Kakade, S. (2001). A Natural Policy Gradient. NeurIPS 2001. Introduce il gradiente naturale per l'ottimizzazione della policy nel reinforcement learning, collegando la geometria di Fisher all'aggiornamento delle policy.

Gibbard, A. (1973). Manipulation of Voting Schemes: A General Result. Econometrica, 41(4), 587-601. Il teorema di Gibbard-Satterthwaite: qualsiasi sistema di votazione surjettivo e non dittatoriale è manipolabile.

Goodhart, C. (1975). Problems of Monetary Management: The U.K. Experience. In Papers in Monetary Economics, Vol. I. Reserve Bank of Australia. La formulazione originaria della legge di Goodhart: quando una misura diventa un obiettivo, cessa di essere una buona misura.

Christiano, P., Leike, J., Brown, T., et al. (2017). Deep Reinforcement Learning from Human Preferences. NeurIPS 2017. Il paper fondativo del RLHF. Introduce l'apprendimento della funzione di ricompensa da comparazioni umane.

Silvano Salvador